

**Dissimilarity Learning under Noise: Classical
Multidimensional Scaling and Randomer Forests**

by

Gongkai Li

A dissertation submitted to The Johns Hopkins University in conformity with the
requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2019

© Gongkai Li 2019

All rights reserved

Abstract

This dissertation mainly concerns with learning from dissimilarities. That is, instead of features, we only observe dissimilarities between pairs of objects for inference. This dissertation is composed with three parts regarding using classical multidimensional scaling under the noise, using Randomer Forest to denoise the datasets and how we can utilize dissimilarity for the task of graph matching.

Classical multidimensional scaling (CMDS) is a widely used method in dimensionality reduction and manifold learning. The method takes in a dissimilarity matrix and outputs a low-dimensional configuration matrix based on a spectral decomposition. In this dissertation, we present three noise models and analyze the resulting configuration matrices, or embeddings. In particular, we show that under each of the three noise models the resulting embedding gives rise to a central limit theorem. We also provide compelling simulations and real data illustrations of these central limit theorems. This perturbation analysis represents a significant advancement over previous results regarding classical multidimensional scaling behavior under randomness.

CMDS is a special case of a more general procedure called Manifold Learning,

ABSTRACT

which is essentially required to achieve quality inferences for modern high-dimensional datasets. Many manifold learning methods have been proposed, each with their own advantages and disadvantages. In this dissertation, building off recent advances in supervised learning, we modify the leading supervised decision forest method to support unsupervised learning, and therefore also nonlinear manifold learning. The key differentiator between our Unsupervised Randomer Forest (URerf) and other manifold learning techniques is that URerF operates on low-dimensional sparse linear combinations of features, rather than either the full observed dimensionality, or one-dimensional marginals. We quantify the efficacy of URerF by computing precision-recall curves relative to the true latent manifold or class label (when it is known). Empirical results on simulated data demonstrate that URerF is robust to high-dimensional noise, where as other methods, such as Isomap and UMAP, quickly deteriorate in such settings.

Finally, in this dissertation, we present a novel approximate graph matching algorithm that incorporates seeded data into the graph matching paradigm. Our Joint Optimization of Fidelity and Commensurability (JOFC) algorithm embeds m graphs into a common Euclidean space where the matching inference task can be performed. Through real and simulated data examples, we demonstrate the versatility of our algorithm in matching graphs with various characteristics—weightedness, directedness, loopiness, many-to-one and many-to-many matchings, and soft seedings.

ABSTRACT

Primary Reader: Carey E. Priebe

Secondary Reader: Minh Tang

Acknowledgments

The past five years in graduate school has been nothing but pure joy and excitement for me. Much of the previous statement is only true because of my advisor, Carey Priebe, without whom this journey will be impossible, let alone enjoyable. I want to thank him for his support and valuable advice at all levels. To Minh Tang and Avanti Athreya, I want to express my uttermost admiration for putting up with me and explain things to me when I am confused (which is most of the time). I also want to thank my co-authors and collaborators, including Nicolas Charon, Vince Lyzinski, Youngser Park, Joshua Vogelstein and Randal Burns, who have all offered their help with patience in discussions over the past few years.

Special thanks to Dan Naiman, Daniel Robinson for their time and suggestions in my Candidacy Exam and Raman Arora and Katia Consani for their time and suggestions in my Graduate Board Exam.

I also want to express my appreciation for the faculty and staff of the Applied Mathematics and Statistics Department at Johns Hopkins University. In particular, my thanks to John Wierman, Edward Scheinerman, Fred Torcaso, Donniell Fishkind,

ACKNOWLEDGMENTS

Tamás Budavári for their kindness advice; and to Kristin Bechtel, Sandy Kirt, and Ann Gibbins for their help. My deepest thanks to my friends: Jordan Yoder, Theo Drivas, Allen Cheng, Heather Pastolic, Wei-Chun Hung, Hsi-Wei Hsieh, Joshua Cape, Chu-Chi Lee, Meghana Madhyastha and Hayden Helm for keeping me sane during the past couple years. Finally, my warmest thanks to my parents, Weiyang Li and Huiping Xu, whose unconditional support made me who I am (well, at least the good part), and also to Cindy, for putting up with me.

Dedication

This thesis is dedicated to my parents and the Schaufelds

Contents

Abstract	ii
Acknowledgments	v
List of Tables	xii
List of Figures	xiii
1 Introduction	1
1.1 Classical Multidimensional Scaling under Perturbation	1
1.2 Randomer Forest for Nearest Neighbour Learning	3
1.3 Seeded Graph Matching Via Joint Optimization of Fidelity and Com- mensurability	6
2 Classical Multidimensional Scaling Under Noise	8
2.1 Review of Classical Multidimensional Scaling	8
2.2 Noise Models and Embedding	9

CONTENTS

2.2.1	Model 1: $\Delta^2 = D^2 + E$	9
2.2.2	Model 2: $\Delta = D + E$	10
2.2.3	Model 3: Matrix Completion	11
2.3	Related Works	11
2.4	Main Theorems	14
2.5	Empirical Results	18
2.5.1	Three Point-mass Simulated Data	18
2.5.2	Shape clustering	21
2.6	Proof of the Theorems	24
2.6.1	Proof of Theorem 2.4.3	24
2.6.2	Adaptation for Theorem 2.4.1 and 2.4.4	43
3	Manifold Denoising with Unsupervised Randomer Forest	49
3.1	Introduction	49
3.2	Related Work	52
3.3	Unsupervised Randomer Forests	54
3.4	Algorithm	55
3.4.1	Overall algorithm	55
3.4.2	Splitting Criteria	56
3.4.2.1	Two-Means Splitting	57
3.4.2.2	Two-Means Splitting with FastBIC	57
3.4.2.3	2-GMM Splitting with BIC	60

CONTENTS

3.4.3	Proximity Matrix Construction	61
3.4.4	Algorithms	61
3.5	Simulations on Synthetic Datasets	61
3.5.1	Synthetic Datasets	66
4	Seeded Graph Matching Via Joint Optimization of Fidelity and Commensurability	69
4.1	Introduction	69
4.2	The classical graph matching problem and the SGM algorithm	71
4.3	Seeded graph matching via JOFC	73
4.3.1	Setup	74
4.3.2	Embedding the seeded graph data	76
4.3.3	Embedding the unseeded vertices	82
4.3.4	Matching the unseeded vertices	84
4.4	Application: Matching zebrafish brains	87
5	Discussion	89
5.1	Discussion on CLT for CMDS	89
5.2	Conjecture: CMDS on Omni Embedding of graphs and Hypothesis Testing	96
5.2.1	Problem Formulation	96
5.2.2	Simulation for the Choice of Scaling Factor	99

CONTENTS

5.2.2.1	Under the Alternative	99
5.2.2.2	Under the Null	102
5.2.3	Mouse Connectome	105
Vita		119

List of Tables

2.1	Empirical average of covariance matrix $\widehat{\Sigma}^{(1)}$, and entry-wise variance (500 simulations).	19
5.1	Sample Covariance and Scaled Sample Covariance Matrix	103

List of Figures

2.1	Simulation results for $n=50, 100, 500$ and 1000 points, as described in Section 2.5.1. The blue ellipses are the 95% level curves of the empirical covariance matrix, and the blue dots are the empirical centers for three classes. The black dots are the true positions of x_1, x_2 and x_3 , and the black ellipses are the 95% level curve for the theoretical covariance matrices as in Theorem 2.4.3. Note that the blue and black centers and ellipses coincide for large n	20
2.2	Examples from the Kimia Dataset.	21
2.3	Noisy versions of examples from the Kimia Dataset.	22
2.4	Pairs plot of CMDS into \mathbb{R}^3 for the noisy curves. Colors correspond to the different classes (blue for bottle, red for bone, and orange for wrench). The position of the nine template curves in the configuration are highlighted with large black dots.	23
3.1	Synthetic Datasets	67
3.2	Geodesic Precision at $K=50$ with varying noise dimension d'	68
4.1	For graphs G_1 and G_2 , the matching $M_{1,2} \subset V_1 \times V_2$ is given via the gray arrows and is formally defined as $M^{1,2} = \{(1, 1), (1, 2), (1, 3), (2, 5), (3, 4), (4, 6), (5, 8)\}$. Note that vertex 1 in G_1 is matched to three vertices in G_2 , namely vertices 1, 2, and 3. Also vertex 7 in G_2 is not matched to any vertex in G_1	75

LIST OF FIGURES

4.2	The pairwise distance heat map of the unseeded vertices from 3 graphs, at time $t=11$, $t=12$, and $t=13$. The darker colors corresponding to smaller distances and the X axis represents the unseeded vertices 301–469 in one graph and Y axis represents the unseeded vertices 469–301 in another graph. Since we know the true matchings between all vertices, we should ideally see a dark line from upper left to lower right between $t=11$ and $t=12$, which can be readily seen. However, since there is an anomaly at $t = 13$, the unseeded vertices at $t=13$ will not be matched correctly, which also can be seen.	88
5.1	Simulation of CMDS with heteroscedastic noise \tilde{E} . The black dots are the true positions for the three points. The blue dots are the empirical means and the blue ellipses are the 95% level curve of the empirical covariance matrix. Note that \tilde{E} used in this simulation is of the same order for the off-diagonal blocks as that used in Figure 2.1. NB: there is asymptotic bias.	93
5.2	Simulation of MDS using raw stress criterion for $n=50$, 100, 500 and 1000 points. The black dots are the true positions of x_1 , x_2 and x_3 , the blue dots are the empirical mean of the simulation and the blue ellipses are the 95% level curve of the empirical covariance matrix. . .	94
5.3	$n=10$, $m=500$	101
5.4	$n=10$, $m=1000$	101
5.5	$n=100$, $m=100$	102
5.6	$n=500$, $m=20$	103
5.7	$n=1000$, $m=20$	104
5.8	$n=2000$, $m=20$	104
5.9	Embedding Dimension Selection of OMNI Matrix.	106
5.10	Dissimilarity Matrix Between Twenty-four Graphs.	107
5.11	CMDS Embedding of the Dissimilarity Matrix.	108

Chapter 1

Introduction

This dissertation is mainly comprised of three parts, each represents a different aspect of how one can utilize dissimilarity measures, possibility under noise, to perform statistical inference. We briefly introduce each of these three topics in this chapter.

1.1 Classical Multidimensional Scaling under Perturbation

Inference based on dissimilarities is of fundamental importance in statistics, data mining and machine learning Pekalska and Duin [2005], with applications ranging from neuroscience Vogelstein et al. [2014] to psychology Carroll and Chang [1970] to economics Machado and Mata [2015]. In each of these fields, rather than directly

CHAPTER 1. INTRODUCTION

observing the feature values of the objects, often we observe only the dissimilarities or “distances” between pairs of objects (inter-point distances). A common approach to dimensionality reduction and subsequent inference problems involving dissimilarities is to embed the observed distances into some (usually Euclidean) space to recover a configuration that faithfully preserves observed distances, and then proceed to perform inference based on the resulting configuration Borg and Groenen [2005], Cox and Cox [2008], de Leeuw and Heiser [1982], Torgerson [1952]. The popular classical multidimensional scaling (CMDS) dimensionality reduction method provides an example of such an embedding scheme into Euclidean space, in which we have readily available tools to perform statistical inference. Furthermore, CMDS also forms the basis for several other more recent approaches to nonlinear dimension reduction and manifold learning Chen and Buja [2009], Schölkopf et al. [1998], such as Isomap Tenenbaum et al. [2000] and Random Forest manifold learning Criminisi and Shotton [2013] among others.

Although widely used, the behavior of CMDS under randomness remains largely unexplored. Several recent papers have highlighted this omission. Zhang et al. [2016] write “Despite the popularity of multi-dimensional scaling, very little is known about to what extent the distances between the embedded points could faithfully reflect the true pairwise distances when observed with noise.”; Fan et al. [2018] write “[W]e are not aware of any statistical results measuring the performance of MDS under randomness, such as perturbation analysis when the objects are sampled from a probabilistic

model.” and Peterfreund and Gavish [2018] write “To the best of our knowledge, the literature does not offer a systematic treatment on the influence of ambient noise on MDS embedding quality.” This paper addresses this acknowledged gap in the literature.

1.2 Randomer Forest for Nearest Neighbour Learning

The accuracy, scalability, and applicability of many machine learning algorithms is currently impeded by the high-dimensional and large-scale nature of most modern data sets. In particular, the dimensionality, or number of features, of many data sets is often high, often due to noise in the data – each data point is represented as a high-dimensional vector, but only a subset of them actually carries signals for subsequent inference. In other words, the data may live near some unknown low-dimensional manifold embedded in some high-dimensional space. To gain a thorough understanding of the data, it is therefore often necessary to reduce its dimensionality in a way that preserves its underlying structure. *Manifold learning* is a set of tools designed to recover the underlying latent low-dimensional manifold structures of high-dimensional data.

Existing manifold learning methods, however, face a number of challenges. Linear approaches such as principal component analysis (PCA) Pearson [1901], independent

CHAPTER 1. INTRODUCTION

component analysis (ICA) Hyvärinen and Oja [2000], canonical correlation analysis (CCA) Hotelling [1936], multidimensional scaling (MDS) Cox and Cox [2000], CUR decompositions Mahoney and Drineas [2009], and Fisher’s linear discriminant analysis (LDA), have been widely applied and useful in many domains, but make fairly strong assumptions of about a linear structure underlying a data set. To mitigate these issues a number of methods that can be characterized as kernel PCA methods were devised Schölkopf et al. [1997], including Isomap Tenenbaum et al. [2000], Laplacian eigenmaps Belkin and Niyogi [2002], maximum variance unfolding Weinberger and Saul [2006]. These approaches are quite fragile to algorithm parameters, and typically require $\mathcal{O}(n^3)$ operations for n samples, which is prohibitively computationally expensive for many datasets. Methods based on exact nearest neighbors, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) Maaten and Hinton [2008], and Uniform Manifold Approximation and Projection (UMAP) McInnes and Healy [2018] also suffer computationally for large n . Approximate nearest neighbor approaches can mitigate some of these computational issues. For example, Fast Approximate Nearest-Neighbor Matching (FLANN) Muja and Lowe [2014] is a popular algorithm for nearest-neighbor detection in high-dimensional data sets. But FLANN, like all the above mentioned manifold learning algorithms, always operates on the observed dimensionality of the data. When the true manifold is low-dimensional, and the data are high-dimensional, the additional noise dimensions will be problematic for any of these algorithms.

CHAPTER 1. INTRODUCTION

We there propose an approach that we dub *Unsupervised Randomer Forest* (URerF). Unlike the previously described methods, URerF does not need to compute geodesic distances between pairs of points. Instead, URerF examines local structure by recursively clustering data in a sparse linear subspace of the original data, building on the recently proposed randomer forest algorithm for supervised learning Tomita et al. [2015]. This randomer forest approach allows URerF to separate meaningful structure in the data from the noise dimensions.

Another contribution of this manuscript is a novel method for evaluating manifold learning algorithms. Most existing manuscripts on the topic either embed the data into some low-dimensional space, such as 2D or 3D, and then merely visualize the results. This approach is obviously limited in a number of ways: (1) it is purely qualitative, (2) when the structure is higher dimensional it may be lost, and (3) it relies on an embedding, which introduces additional complications. Other manuscripts compare the results on some subsequent inference task, such as classification. Such an approach is only able to evaluate performance of the manifold learning algorithm composed with a particular subsequent inferential method, but not the manifold learning algorithm itself. We therefore introduce Precision@K, Recall@K, and Precision-Recall curves as quantitative metrics to evaluate manifold learning. The difference between our proposed metrics and standard metrics, is that we do not evaluate nearest neighbors with respect to the high-dimensional observed data, but rather the true low-dimensional latent representations. If a manifold learning does

poorly on this metric, it has no hope to perform well on subsequent tasks. Indeed Precision@k provides a theoretical bound on subsequent classification accuracy Devroye et al. [1997].

1.3 Seeded Graph Matching Via Joint Optimization of Fidelity and Commensurability

Given m graphs, the graph matching problem (GMP) seeks to find a set of correspondences (i.e., “matchings”) between the vertex sets that best preserves similar substructures across the graphs. The graph matching problem has applications across many diverse disciplines including document processing, mathematical biology, network analysis and pattern recognition, to name a few. Unfortunately, no graph matching algorithm is known to be efficient. Indeed, even the easier problem of matching isomorphic simple graphs is of famously unknown complexity (see Garey and Johnson [1979]). Because of its practical applicability, there exist numerous approximate graph matching algorithms in the literature; for an excellent survey of the existing literature, see Conte et al. [2004].

When matching across m graphs, often partial correspondences, or seedings, between the vertices of some pairs of graphs are known. One cutting-edge algorithm for

CHAPTER 1. INTRODUCTION

seeded graph matching, the Seeded Graph Matching (SGM) algorithm of Fishkind et al. [2012] and Lyzinski et al. [2013], leverages the information contained in seeded vertices to efficiently match graphs with thousands of vertices, achieving excellent performance with relatively few seeds. However, as demonstrated in Fishkind et al. [2012], SGM achieves its optimal performance in the case of highly structured simple graphs on identical vertex sets. Although it can be modified to handle directed, weighted, and other non-simple graphs, in the presence of these generalizations the performance of the SGM algorithm deteriorates. Moreover, the algorithm cannot currently handle cases where the graphs have different numbers of vertices or matchings across graphs that are not one-to-one. Often graphs arising from real data contain many of the aforementioned characteristics, and more robust procedures are needed to effectively match these graphs.

Herein we present a new seeded graph matching algorithm derived from the Joint Optimization of Fidelity and Commensurability (JOFC) algorithm of Priebe et al. [2013]. Our algorithm is flexible enough to handle many of the difficulties inherent to real data, while simultaneously not sacrificing too much performance (compared to SGM) when matching across simple graphs.

Chapter 2

Classical Multidimensional Scaling Under Noise

2.1 Review of Classical Multidimensional Scaling

Given an $n \times n$ hollow symmetric dissimilarity matrix D , and an embedding dimension d , we seek $X \in \mathbb{R}^{n \times d}$, where the rows $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ of X represent coordinates of points in \mathbb{R}^d , such that the overall inter-point distances between X_i and X_j are as close as possible to the distances given by the dissimilarity matrix D . For a given matrix H , we shall denote by $H^{(2)} = H \circ H$ the element-wise squaring of the matrix H . Given D , classical multidimensional scaling involves the following

CHAPTER 2. CMDS WITH PERTURBATION

steps:

1. Compute the matrix $B = -\frac{1}{2}PD^{(2)}P$ where $P = I - \mathbf{1}_n \mathbf{1}_n^\top / n$ is the double centering matrix. Here I denotes the $n \times n$ identity matrix and $\mathbf{1}_n = (1, \dots, 1)^\top \in \mathbb{R}^n$.
2. Extract the d largest positive eigenvalues s_1, \dots, s_d of B and the corresponding eigenvectors u_1, \dots, u_d .
3. Let $X = U_B S_B^{1/2} \in \mathbb{R}^{n \times d}$, where $U_B = (u_1, \dots, u_d)$ and $S_B = \text{diag}(s_1, \dots, s_d)$.
Each row of X represents the coordinate of a point in \mathbb{R}^d .

In essence, the procedure minimizes the Strain loss function defined as $L(X) = \|XX^\top - B\|_F$ where $\|\cdot\|_F$ denote the Frobenius norm of a matrix. Furthermore, the resulting configuration X centers all points around the origin, resulting in an inherent issue of identifiability: X is unique only up to an orthogonal transformation. In the following presentation, we will write $X = U_B S_B^{1/2} W$ where W is some orthogonal matrix, for a suitably transformed X .

2.2 Noise Models and Embedding

2.2.1 Model 1: $\Delta^2 = D^2 + E$

In this section we propose three different but related noise models for the matrix of observed dissimilarities. Suppose that we have a latent or unobserved matrix D

CHAPTER 2. CMDS WITH PERTURBATION

of inter-point Euclidean distances between n points in \mathbb{R}^d , i.e. $D_{ij} = \|x_i - x_j\|$. Let $D^{(2)}$ denote the entry-wise square of D and Δ be the observed dissimilarity matrix, such as that measured via a scientific experiment.

The first noise model we consider is $\Delta^{(2)} = D^{(2)} + E$ where we think of $D^{(2)}$ as the signal matrix and E as the noise matrix; see also Zhang et al. [2016]. We shall assume that E satisfies the following conditions:

- (i) $\mathbb{E}[E] = 0$, hence $\mathbb{E}[\Delta^{(2)}] = D^{(2)}$.
- (ii) The matrix E is hollow and symmetric.
- (iii) The entries E_{ij} are independent and $\text{Var}(E_{ij}) = \sigma^2$.
- (iv) There exists a finite constant C such that E_{ij} follows a sub-Gaussian distribution with variance proxy C for all i, j , i.e., $\mathbb{P}[E_{ij} \geq t] \leq 2 \exp(-t^2/(2C))$ for all i, j .

2.2.2 Model 2: $\Delta = D + E$

The second error model we consider is $\Delta = D + E$. We once more require that the random matrix E satisfies conditions (i) to (iv) identical to that in the model $\Delta^{(2)} = D^{(2)} + E$ along with a constant third and fourth moment conditions, i.e., there exists finite constants γ and ξ such that (v) $\mathbb{E}[E_{ij}^3] \equiv \gamma$ and $\mathbb{E}[E_{ij}^4] \equiv \xi$ for all i, j .

2.2.3 Model 3: Matrix Completion

Finally, we consider a noise model where only a fraction of the entries of D are observed. More specifically, for a given $q \in [0, 1]$ let Δ be such that for $i < j$, with probability q we observe $\Delta_{ij} = D_{ij}$ and with probability $1 - q$, Δ_{ij} is unobserved in which case we set $\Delta_{ij} = 0$. We then have $\Delta = D + E$ where E_{ij} is distributed as $(-D_{ij}) \times \text{Bernoulli}(1 - q)$. Furthermore, $\mathbb{E}[\Delta] = q \cdot D$ and $E[\Delta^{(2)}] = q \cdot D^{(2)}$. This model is motivated by the widely-studied problems of distance matrix completion and sensor localization; see e.g., Alfakih et al. [1999], Chatterjee [2015], Javanmard and Montanari [2013], Patwari et al. [2005].

For each of the above noise models, we shall apply classical multidimensional scaling to the observed Δ to obtain a configuration matrix \hat{X} whose rows are the estimate of the latent, unobserved $X = [x_1, \dots, x_n]^\top$. A natural question that arises is how the added noise affects the embedding configuration. That is, what is the relationship between the configuration X and \hat{X} obtained from classical multidimensional scaling of D and Δ ?

2.3 Related Works

The problem of recovering an Euclidean distance matrix from noisy or imperfect observations of pairwise dissimilarity scores arises naturally in many different contexts. For example, in Zhang et al. [2016], the authors considered the model

CHAPTER 2. CMDS WITH PERTURBATION

$\Delta^{(2)} = D^{(2)} + E$ along with the estimator

$$\widehat{D}^{(2)} = \arg \max_{M \in \mathcal{D}_n^{(2)}} \left\{ \frac{1}{2} \|\Delta^{(2)} - M\|_F^2 + \lambda_n \text{trace} \left(-\frac{1}{2} P M P \right) \right\}$$

for $D^{(2)}$. Here $\mathcal{D}_n^{(2)}$ is the set of $n \times n$ *squared* Euclidean distance matrix and λ_n is a tuning parameter. Corollary 6 in Zhang et al. [2016] states that under suitable model on E , with probability approaching to one we have

$$\|\widehat{D}^2 - D^2\|_F^2 \leq 36n\sigma^2(r+1) \quad (2.1)$$

where σ is the variance of the noise and r is the rank of D^2 . In this paper we obtain, as a corollary of ours results, a bound of the same order on $\|\widehat{D}^2 - D^2\|_F^2$. Furthermore, our central limit theorem on the configuration matrix \widehat{X} provides a more refined limiting result, albeit one of a different flavor from Eq. (2.1).

On the other hand, completing a distance matrix with missing entries has been a popular problem in the engineering and social sciences; see, for example, Alfakih et al. [1999], Bakonyi and Johnson [1995], Singer [2008], Spence and Domoney [1974]. Distance matrix completion is closely related to multidimensional scaling [Borg and Groenen, 2005, Chatterjee, 2015, Javanmard and Montanari, 2013, Oh et al., 2010]. Especially noteworthy is Theorem 2.5 of Chatterjee [2015], where the author established an upper bound for the mean squared error on the estimator \widetilde{M} for a general distance matrix M . More specifically, let (K, d) be a compact metric space

CHAPTER 2. CMDS WITH PERTURBATION

and x_1, \dots, x_n be n arbitrary points in K . Let M be the $n \times n$ matrix whose ij -entry is $d(x_i, x_j)$. Let $\epsilon > 0$ be such that $q \geq n^{-1+\epsilon}$. For a given $\delta > 0$, let $N(\delta)$ be the covering number of K using balls of radius ϵ with respect to the metric d . Then there exists an estimator \widetilde{M} obtained by truncating the singular value decomposition of M such that

$$\text{MSE}(\widetilde{M}) \leq C \inf_{\delta > 0} \min \left\{ \frac{\delta + \sqrt{N(\delta/4)/n}}{\sqrt{q}}, 1 \right\} + C(\epsilon)e^{-ncq}$$

where c and C are constants depending on the truncation level η for the singular values of M and $C(\epsilon)$ is a constant depending only on ϵ and η . Of particular interest is the application of this theorem to the Euclidean distance matrix, for which we obtain roughly

$$\text{MSE}(\widetilde{M}) \leq \frac{Cn^{-1/3}}{\sqrt{q}}.$$

Another recent result for the configuration \widehat{X} obtained from the incomplete distance matrix Δ is Theorem 1 of Taghizadeh [2014] which states that, with high probability

$$\|\widehat{X} - X\|_F \leq \mathcal{O}\left(\frac{\sqrt{n}}{\sqrt{q}}\right).$$

Our central limit theorem in this paper improves upon both results. It is worth mentioning that the Euclidean distance matrix completion problem can also be viewed from an optimization point of view. See Tasissa and Lai [2018] for a review of such approaches.

2.4 Main Theorems

Recall that a random variable X is sub-Gaussian if

$$\mathbb{P}[|X| > t] \leq 2e^{-\frac{t^2}{K^2}}$$

for some constant K and for all $t \geq 0$. Associated with a sub-Gaussian random variable X is a Orlicz norm defined as

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp(\frac{X^2}{t^2}) \leq 2\}.$$

A random vector X in \mathbb{R}^n is called sub-Gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-Gaussian random variables for all $x \in \mathbb{R}^n$, and the corresponding sub-Gaussian norm of X is defined as

$$\|X\|_{\psi_2} = \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_2}.$$

We now present central limit theorems for the rows of the classical multidimensional scaling configuration \hat{X} for the three noise models in § 2.2.1. Intuitively speaking, the theorems established that the rows of \hat{X} , after some orthogonal transformation, is approximately normally distributed around the rows of X . Furthermore, the covariance matrix will depend on the noise model and the true distribution of the points in the underlying space and are substantially different between the three

CHAPTER 2. CMDS WITH PERTURBATION

noise models considered. In particular, the covariance matrix for the noise model $\Delta^2 = D^2 + E$ in Theorem 2.4.1 depends only on the variance σ^2 of the noise E_{ij} . This is in contrast with the covariance matrices of the model $\Delta = D + E$ and the model $\mathbb{E}[\Delta] = qD$ in Theorem 2.4.3 and Theorem 2.4.4, both of which depend also on the underlying true distances D_{ij} . The machinery involved in proving these results are by and large the same and we refer the reader to the Appendix for a sketch of the proof. For ease of exposition, we denote by $(A)_i$ the i -th row of a matrix.

Theorem 2.4.1 (central limit theorem for $\Delta^2 = D^2 + E$) *Let Z_1, \dots, Z_n be independent and identically distributed according to a multivariate sub-Gaussian distribution F on \mathbb{R}^d . Let D be the Euclidean distance matrix generated by the Z_k 's, i.e. $D_{ij} = \|Z_i - Z_j\|$. Let $\Delta^2 = D^2 + E$ where the noise matrix E satisfy the conditions*

1. $\mathbb{E}[E] = \mathbf{0}$,
2. E is hollow (0 on the diagonal) and symmetric,
3. the entries E_{ij} are independent for $i \leq j$ with $\text{Var}[E_{ij}] \equiv \sigma^2$, and
4. each E_{ij} follows a sub-Gaussian distribution.

Denote by \hat{X}_n the classical multidimensional scaling embedding configurations of Δ into \mathbb{R}^d . There exists a sequence of $d \times d$ orthogonal matrices $\{W_n\}_{n=1}^\infty$ such that for any $\alpha \in \mathbb{R}^d$ and any fixed row index i , we have

$$\lim_{n \rightarrow \infty} \mathbb{P}\{n^{1/2}[(\hat{X}_n W_n)_i - (Z_i - \bar{Z})] \leq \alpha\} = \Phi(\alpha, \Sigma)$$

CHAPTER 2. CMDS WITH PERTURBATION

where \bar{Z} is the mean of Z_k 's and $\Phi(\alpha, \Sigma)$ denotes the cumulative distribution function of a multivariate Gaussian with mean 0 and covariance matrix Σ , evaluated at α . Here $\Sigma = \frac{\sigma^2}{4} \Xi^{-1}$ where $\Xi = \text{cov}(Z_k) \in \mathbb{R}^{d \times d}$.

Remark 2.4.2 We can relax the common variance requirement (iii) in Theorem 2.4.1.

Let $\text{Var}(E_{ij}) = \sigma_{ij}^2$ and suppose that, for a fixed i , the collection $(D_{ij}^2 - \Delta_{ij}^2)(Z_j - \mathbb{E}[Z_j])$ for $j \neq i$ satisfies the conditions for the Lindeberg-Feller central limit theorem. Let $\Sigma_i = n^{-1} \sum_j \sigma_{ij}^2 \text{cov}(Z_k)$. We obtain the following variant of Theorem 2.4.1:

$$n^{1/2} \Sigma_i^{-\frac{1}{2}} ((\widehat{X}_n W_n)_i - (Z_i - \bar{Z})) \rightarrow \mathcal{N}(0, I).$$

Theorem 2.4.3 (Central Limit Theorem for $\Delta = D + E$) Let Z_1, \dots, Z_n be independent and identically distributed according to a multivariate sub-Gaussian distribution F on \mathbb{R}^d . Let D be the Euclidean distance matrix generated by the Z_k 's, i.e. $D_{ij} = \|Z_i - Z_j\|$. Let $\Delta = D + E$ and suppose that the noise matrix E satisfy, in addition to the conditions in Theorem 2.4.1, the condition (5) $\mathbb{E}[E_{ij}^3] \equiv \gamma$ and $\mathbb{E}[E_{ij}^4] \equiv \xi$. Denote by \widehat{X}_n the classical multidimensional embedding of Δ into \mathbb{R}^d . There exists a sequence of $d \times d$ orthogonal matrices $\{W_n\}_{n=1}^\infty$ such that for any $\alpha \in \mathbb{R}^d$ and any fixed row index i ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\{n^{1/2} ((\widehat{X}_n W_n)_i - (Z_i - \bar{Z})) \leq \alpha\} = \int \Phi(\alpha, \Sigma(z)) dF(z)$$

CHAPTER 2. CMDS WITH PERTURBATION

where \bar{Z} is the mean of Z_k 's and $\Phi(\alpha, \Sigma)$ denotes the cumulative distribution function of a multivariate Gaussian with mean 0 and covariance matrix Σ , evaluated at α . Here $\Sigma(z) = \Xi^{-1}\tilde{\Sigma}(z)\Xi^{-1}$ where $\Xi = \text{cov}(Z_i) \in \mathbb{R}^{d \times d}$ and, with $\mu = \mathbb{E}[Z_i] \in \mathbb{R}^d$,

$$\tilde{\Sigma}(z) = \mathbb{E}_{Z_k} \left[(\sigma^2 \|z - Z_k\|^2 + \gamma \|z - Z_k\| + \frac{1}{4}\xi - \frac{\sigma^4}{4})(Z_k - \mu)(Z_k - \mu)^\top \right]$$

is a covariance matrix depending on z .

Theorem 2.4.4 (Central Limit Theorem for $\Delta = D$ with missing entries) *Let Z_1, \dots, Z_n be independent and identically distributed according to a multivariate sub-Gaussian distribution F on \mathbb{R}^d . Let D be the Euclidean distance matrix generated by the Z_i 's, i.e. $D_{ij} = \|Z_i - Z_j\|$. Suppose that with probability $q_n \in [0, 1]$ we observe the distance D_{ij} and with probability $1 - q_n$ it is missing, i.e., $\Delta = D + E$ where $E_{ij} = (-D_{ij}) \times \text{Bernoulli}(1 - q_n)$. Denote by \hat{X}_n the classical multidimensional embedding of Δ into \mathbb{R}^d . Then there exists a sequence of $d \times d$ orthogonal matrices $\{W_n\}_{n=1}^\infty$ such that if $nq_n = \omega(\log^4 n)$, then for any $\alpha \in \mathbb{R}^d$ and any fixed row index i ,*

$$\lim_{n \rightarrow \infty} \mathbb{P}\{n^{1/2}[(\hat{X}_n W_n)_i - q_n^{1/2}(Z_i - \bar{Z})] \leq \alpha\} = \int \Phi(\alpha, \Sigma(z)) dF(z)$$

where \bar{Z} is the mean of Z_i 's and $\Phi(\alpha, \Sigma)$ denotes the CDF of a multivariate Gaussian with mean 0 and covariance matrix Σ , evaluated at α . Here $\Sigma(z) = \Xi^{-1}\tilde{\Sigma}(z)\Xi^{-1}$,

CHAPTER 2. CMDS WITH PERTURBATION

$\Xi = \text{cov}(Z_i) \in \mathbb{R}^{d \times d}$ and with $\mu = \mathbb{E}[Z_i] \in \mathbb{R}^d$,

$$\tilde{\Sigma}(z) = \frac{1 - q_n}{4} \times \mathbb{E}_{Z_k} \left[\|z - Z_k\|^4 (Z_k - \mu)(Z_k - \mu)^\top \right]$$

is a covariance matrix depending on z .

Remark 2.4.5 *We emphasize that, in the statement of Theorem 2.4.4, $(\widehat{X}_n W_n)_i$ is centered around $q_n^{1/2}(Z_i - \bar{Z})$ and not around $Z_i - \bar{Z}$. Therefore, unless q_n is known or that an identifiability condition is specified, the classical multidimensional scaling configuration \widehat{X}_n will only recovers an estimate of $Z - 1_n \bar{Z}$ up to an orthogonal transformation W_n and a scaling factor $q_n^{1/2}$.*

2.5 Empirical Results

2.5.1 Three Point-mass Simulated Data

As a simple illustration of our central limit theorem, we embed noisy Euclidean distances obtained from n points into \mathbb{R}^2 . For illustrative purpose, we will focus on the error model $\Delta = D + E$ as in Theorem 2.4.3. Experimental results for the other error models are completely analogous. We consider three points $x_1, x_2, x_3 \in \mathbb{R}^2$ for which the inter-point distances are 3, 4 and 5 (these three points form a right triangle) and generate $n_k = \pi_k n$ points equal to x_k , $k = 1, 2, 3$, where $\pi = [0.2, 0.3, 0.5]^\top$. The resulting Euclidean inter-point distance matrix D is then subjected to uniform noise,

CHAPTER 2. CMDS WITH PERTURBATION

	$n=50$	$n=100$	$n=500$	$n=1000$
$\widehat{\Sigma}^{(1)} :$	$\begin{bmatrix} 14.15 & 0.25 \\ 0.25 & 79.07 \end{bmatrix}$	$\begin{bmatrix} 13.67 & -0.79 \\ -0.79 & 98.96 \end{bmatrix}$	$\begin{bmatrix} 13.65 & -2.34 \\ -2.34 & 41.02 \end{bmatrix}$	$\begin{bmatrix} 13.63 & -2.70 \\ -2.70 & 31.76 \end{bmatrix}$
$\text{Var} \begin{bmatrix} \widehat{\Sigma}_{11}^{(1)} \\ \widehat{\Sigma}_{12}^{(1)} \\ \widehat{\Sigma}_{22}^{(1)} \end{bmatrix} :$	$\begin{bmatrix} 41.25 \\ 113.31 \\ 829.52 \end{bmatrix}$	$\begin{bmatrix} 19.29 \\ 68.06 \\ 984.45 \end{bmatrix}$	$\begin{bmatrix} 3.67 \\ 7.87 \\ 31.71 \end{bmatrix}$	$\begin{bmatrix} 1.71 \\ 3.25 \\ 11.08 \end{bmatrix}$

Table 2.1: Empirical average of covariance matrix $\widehat{\Sigma}^{(1)}$, and entry-wise variance (500 simulations).

yielding $\Delta = D + E$ where $E_{ij} \stackrel{i.i.d.}{\sim} \text{Uniform}(-4, +4)$ for $i < j$ and $E_{ij} = E_{ji}$. For this setting, our central limit theorem for the classical multidimensional embedding of Δ into \mathbb{R}^2 yields class-conditional Gaussians. For $n \in \{50, 100, 1000\}$, Figure 2.1 compares, for one realization, the theoretical vs. estimated means and covariances matrices (95% level curves). Table 2.1 shows the empirical covariance matrix for one of the point masses, $\widehat{\Sigma}^{(1)}$, behaving in accordance with Theorem 2.4.3.

Table 2.1 investigates the empirical covariance matrix for one of the point masses, and its entry-wise variance, as a function of n . The theoretical covariance matrix is

$$\Sigma^{(1)} = \begin{bmatrix} 13.56 & -3.06 \\ -3.06 & 22.65 \end{bmatrix}$$

Remark 2.5.1 *In this simulation we relax the requirement that the entries of Δ should be nonnegative in order to illustrate the phenomenon of decreasing covariance with increasing n .*

CHAPTER 2. CMDS WITH PERTURBATION

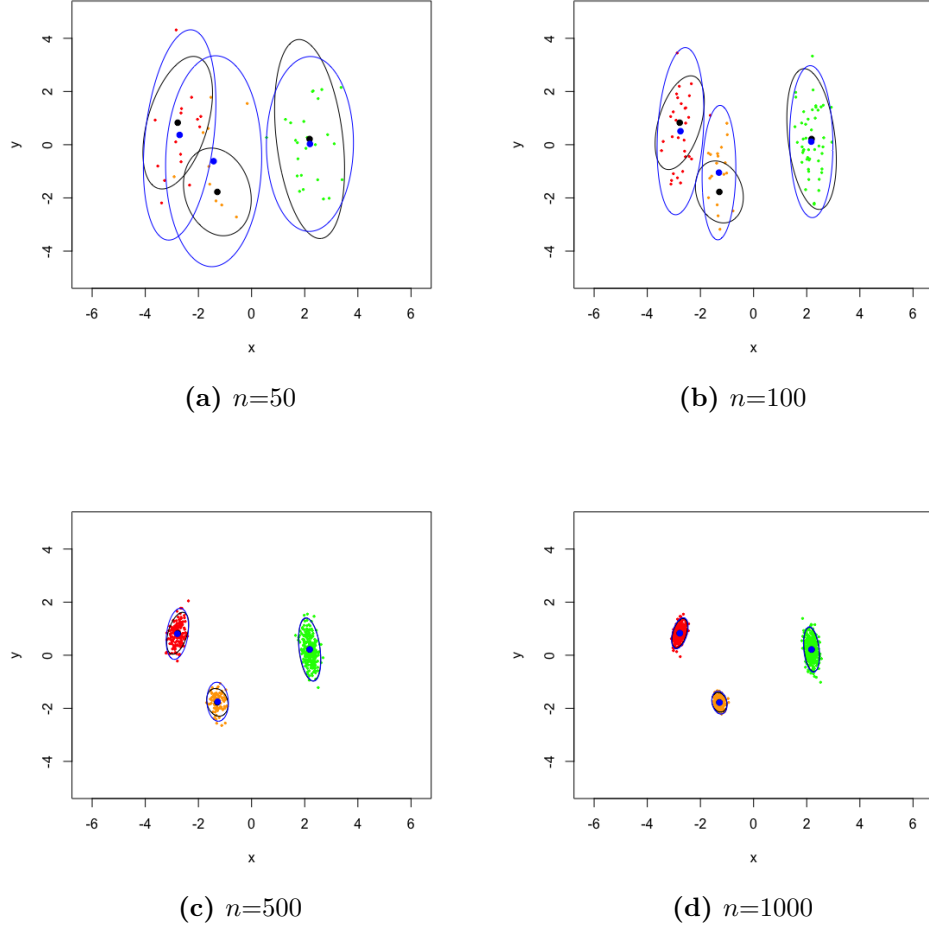


Figure 2.1: Simulation results for $n=50$, 100, 500 and 1000 points, as described in Section 2.5.1. The blue ellipses are the 95% level curves of the empirical covariance matrix, and the blue dots are the empirical centers for three classes. The black dots are the true positions of x_1 , x_2 and x_3 , and the black ellipses are the 95% level curve for the theoretical covariance matrices as in Theorem 2.4.3. Note that the blue and black centers and ellipses coincide for large n .

2.5.2 Shape clustering

As a second illustration of the effect of noise on CMDS, we examine a more involved clustering experiment in the (non-Euclidean) shape space of closed curves. In this experiment, we consider boundary curves obtained from silhouettes of the Kimia shape database. Specifically, we restrict attention to three predefined classes of objects (bottle, bone, and wrench) and take from each class three different examples of shapes all given by planar closed polygonal curves representing the objects' outline. Figure 2.2 shows one instance for each of the bottle, bone, and wrench class. A database of noisy curves is then created as follows: for each of the nine template shapes, we generate 100 noisy realizations in which vertices of the curve are moved along the curve's normal vectors with random distances drawn from independent Gaussian distributions at each vertex. This results in a total of 900 noisy versions of the initial curves such as the ones displayed in Figure 2.3.

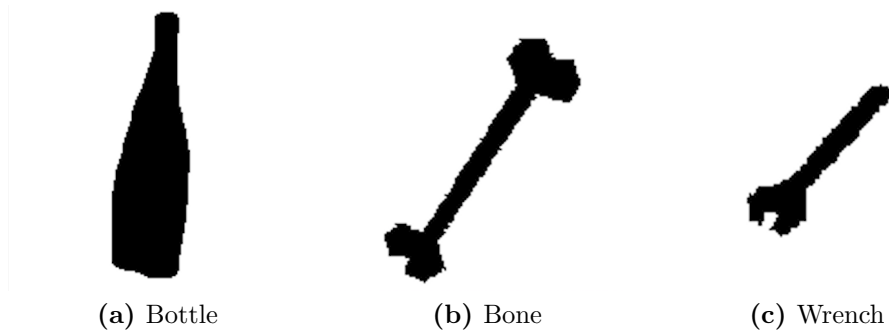


Figure 2.2: Examples from the Kimia Dataset.

We then compute the pairwise distance matrix between all the curves (including the noiseless templates) based on a shape distance which was introduced in Glaunès

CHAPTER 2. CMDS WITH PERTURBATION

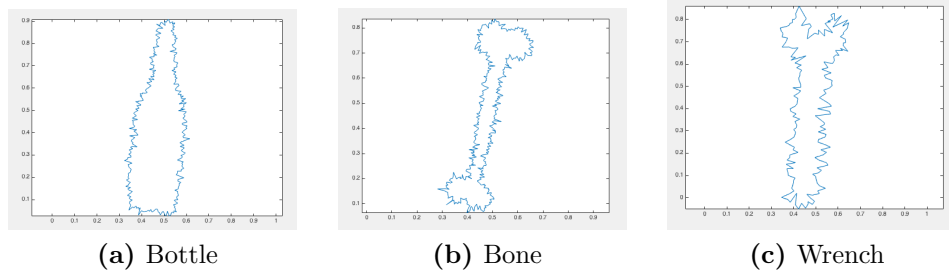


Figure 2.3: Noisy versions of examples from the Kimia Dataset.

et al. [2008] and later extended in the work of Kaltenmark et al. [2017]. This type of metric is based on the representation of shapes in a particular distribution space called currents, see Kaltenmark et al. [2017] for details. In our context, this metric offers several advantages: (i) the distance is completely geometrical in the sense that it is independent of the sampling of the curves and does not rely on predefined pointwise correspondences between vertices; (ii) it has an intrinsic smoothing effect that provides robustness to noise to a certain degree; (iii) it can be computed in closed form with minimal computational time which is critical given the large number of pairwise distances to evaluate. In this setting, we can view the resulting distance matrix as a perturbation of the ideal distances between the 9 template curves, which fits into the generic framework of our model. (Note that we leave aside the issue of checking the technical assumptions on the matrix E , which may be quite involved for this noise model and distance.)

We proceed to perform CMDS on this distance matrix. A scree plot investigation shows that an appropriate embedding dimension here is $\hat{d} = 3$ (the top three eigenvalues are 2.20, 0.68, 0.06 with the fourth $\ll 0.01$). The resulting embedding config-

CHAPTER 2. CMDS WITH PERTURBATION

uration is shown in Figure 2.4. This configuration exhibits nine fairly well-separated clusters roughly centered around the position of each of the noiseless template curves. Those, in turn, form 3 ‘super-clusters’ consistent with the classes. Furthermore, the ellipsoidal shape of each cluster suggests that the configuration approximately follows a Gaussian distribution.

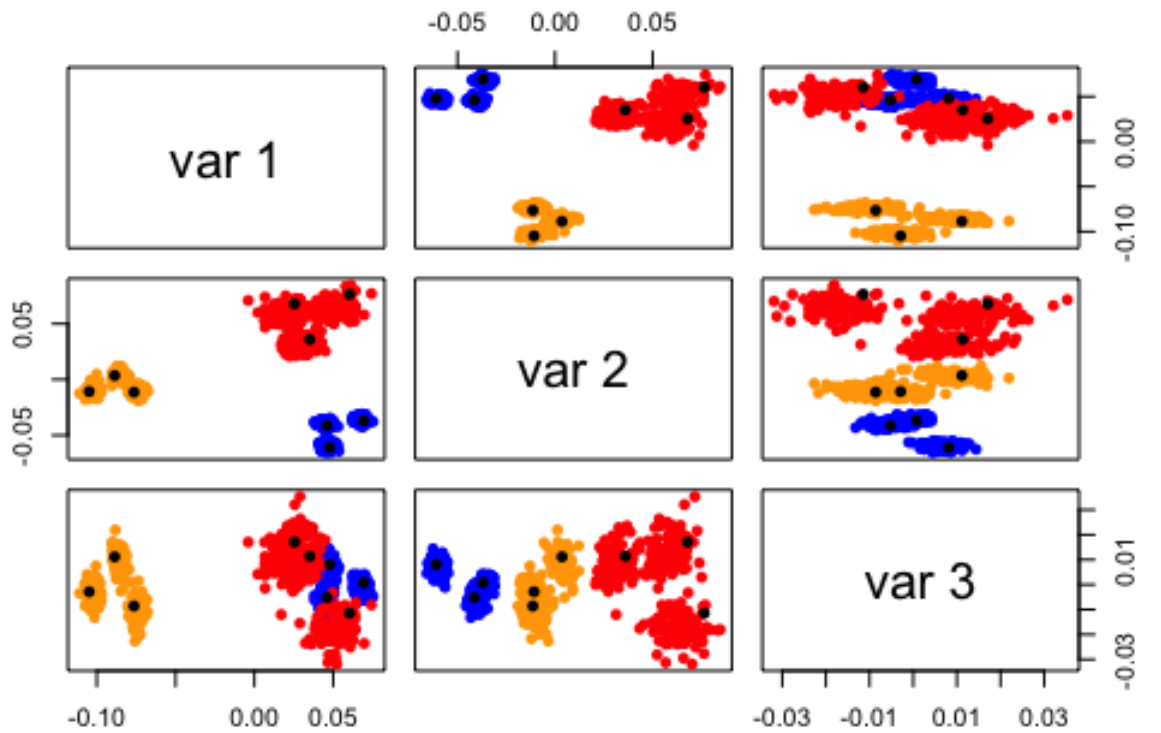


Figure 2.4: Pairs plot of CMDS into \mathbb{R}^3 for the noisy curves. Colors correspond to the different classes (blue for bottle, red for bone, and orange for wrench). The position of the nine template curves in the configuration are highlighted with large black dots.

While these preliminary shape clustering results are obtained with a specific and simple distance on the space of curves, future work will investigate whether similar

properties hold with different, more elaborate metrics and/or geometric noise models. The central limit theorem derived here could then constitute a useful theoretical tool to evaluate the discriminating power of shape clustering methods based on CMDS.

2.6 Proof of the Theorems

2.6.1 Proof of Theorem 2.4.3

We proceed to give a complete proof for Theorem 2.4.3. Theorems 2.4.1 and 2.4.4 will have different covariance matrix structures than what is given in Lemma 2.6.3 and will be dealt with later.

Given a matrix A , we denote by $\|A\|$ and $\|A\|_F$ its spectral and Frobenius norm, respectively. We will utilize the following observation repeatedly in our presentation.

Observation 2.6.1 *Let A and B be matrices of appropriate dimensions. Then*

$$\|AB\|_F = \|B^\top A^\top\|_F \leq \min\{\|A\| \times \|B\|_F, \|B\| \times \|A\|_F\}.$$

We remind our readers the following notations for the subsequent presentation. Recall that

$$B = -\frac{1}{2}PD^{(2)}P$$

CHAPTER 2. CMDS WITH PERTURBATION

and

$$\widehat{B} = -\frac{1}{2}P\Delta^{(2)}P$$

are the double centering of $D^{(2)}$ and $\Delta^{(2)}$, respectively. If $D^{(2)}$ is a Euclidean distance matrix whose elements are $D_{ij} = \|Z_i - Z_j\|$, then

$$B = PZZ^\top P.$$

In particular,

$$U_B S_B^{1/2} = PZ\widetilde{W}$$

for some $d \times d$ orthogonal matrix \widetilde{W} . The i th row of $U_B S_B^{1/2}$ is then $\widetilde{W}_n^\top (Z_i - \bar{Z})$.

Now let W^* be the orthogonal matrix satisfying

$$W^* = \arg \min_W \|U_B^\top U_{\widehat{B}} - W\|_F.$$

The following lemma provides a decomposition for $\widehat{X} - U_B S_B^{1/2} W^*$ into a sum of several matrices.

Lemma 2.6.2 *Let W^* be the orthogonal matrix satisfying*

$$W^* = \arg \min_W \|U_B^\top U_{\widehat{B}} - W\|.$$

CHAPTER 2. CMDS WITH PERTURBATION

Then

$$\widehat{X} - U_B S_B^{1/2} W^* = (\widehat{B} - B) U_B S_B^{-1/2} W^* \quad (2.2)$$

$$- (\widehat{B} - B) U_B (S_B^{-1/2} W^* - W^* S_{\widehat{B}}^{-1/2}) - U_B U_B^\top (\widehat{B} - B) U_B W^* S_{\widehat{B}}^{-1/2} \quad (2.3)$$

$$+ (I - U_B U_B^\top) (\widehat{B} - B) (U_{\widehat{B}} - U_B W^*) S_{\widehat{B}}^{-1/2} \quad (2.4)$$

$$+ U_B (U_B^\top U_{\widehat{B}} - W^*) S_{\widehat{B}}^{1/2} + U_B (W^* S_{\widehat{B}}^{1/2} - S_B^{1/2} W^*) \quad (2.5)$$

Proof: We have

$$\begin{aligned} \widehat{X} - U_B S_B^{1/2} W^* &= U_{\widehat{B}} S_{\widehat{B}}^{1/2} - U_B W^* S_{\widehat{B}}^{1/2} + U_B (W^* S_{\widehat{B}}^{1/2} - S_B^{1/2} W^*) \\ &= U_{\widehat{B}} S_{\widehat{B}}^{1/2} - U_B U_B^\top U_{\widehat{B}} S_{\widehat{B}}^{1/2} + U_B U_B^\top U_{\widehat{B}} S_{\widehat{B}}^{1/2} - U_B W^* S_{\widehat{B}}^{1/2} \\ &\quad + U_B (W^* S_{\widehat{B}}^{1/2} - S_B^{1/2} W^*) \\ &= (I - U_B U_B^\top) \widehat{B} U_{\widehat{B}} S_{\widehat{B}}^{-1/2} + U_B (U_B^\top U_{\widehat{B}} - W^*) S_{\widehat{B}}^{1/2} + U_B (W^* S_{\widehat{B}}^{1/2} - S_B^{1/2} W^*) \\ &= (I - U_B U_B^\top) (\widehat{B} - B) U_{\widehat{B}} S_{\widehat{B}}^{-1/2} \\ &\quad + U_B (U_B^\top U_{\widehat{B}} - W^*) S_{\widehat{B}}^{1/2} + U_B (W^* S_{\widehat{B}}^{1/2} - S_B^{1/2} W^*). \end{aligned}$$

We used the facts $U_B U_B^\top B = B$ and $U_{\widehat{B}} S_{\widehat{B}}^{1/2} = \widehat{B} U_{\widehat{B}} S_{\widehat{B}}^{-1/2}$ in the above equalities.

The last two terms of the above display is Eq. (2.5). Denote

$$R := (I - U_B U_B^\top) (\widehat{B} - B) U_{\widehat{B}} S_{\widehat{B}}^{-1/2},$$

CHAPTER 2. CMDS WITH PERTURBATION

we then have

$$\begin{aligned}
R &= (I - U_B U_B^\top)(\widehat{B} - B)(U_B W^* + U_{\widehat{B}} - U_B W^*) S_{\widehat{B}}^{-1/2} \\
&= (\widehat{B} - B) U_B W^* S_{\widehat{B}}^{-1/2} - U_B U_B^\top (\widehat{B} - B) U_B W^* S_{\widehat{B}}^{-1/2} \\
&\quad + (I - U_B U_B^\top)(\widehat{B} - B)(U_{\widehat{B}} - U_B W^*) S_{\widehat{B}}^{-1/2} \\
&= (\widehat{B} - B) U_B S_B^{-1/2} W^* - (\widehat{B} - B) U_B (S_B^{-1/2} W^* - W^* S_{\widehat{B}}^{-1/2}) \\
&\quad - U_B U_B^\top (\widehat{B} - B) U_B W^* S_{\widehat{B}}^{-1/2} + (I - U_B U_B^\top)(\widehat{B} - B)(U_{\widehat{B}} - U_B W^*) S_{\widehat{B}}^{-1/2}
\end{aligned}$$

The four terms in the above display are identical to that in Eq. (2.2) through Eq. (2.4).

■

Lemma 2.6.2 implies

$$\widehat{X} W^{*\top} \widetilde{W}_n - U_B S_B^{1/2} \widetilde{W}_n = \widehat{X} W^{*\top} \widetilde{W}_n - PZ = (\widehat{B} - B) U_B S_B^{-1/2} \widetilde{W}_n + R_n \widetilde{W}_n$$

where R_n are the matrices in Eq. (2.3) through Eq. (2.5). The essential term is

$$(\widehat{B} - B) U_B S_B^{-1/2} \widetilde{W}_n.$$

We analyzed the rows of this matrix in Lemma 2.6.3 where we show that they converge to multivariate normals. Meanwhile, Lemma 2.6.4 shows that the rows of the matrices R_n , when scaled by $n^{1/2}$, converge to 0 in probability. Combining these results yield Theorem 2. A few minor changes to the covariance computation in the proof of

CHAPTER 2. CMDS WITH PERTURBATION

Lemma 2.6.3 also yield Theorem 2.4.1 and Theorem 2.4.4.

Lemma 2.6.3 *Let Z_1, \dots, Z_n be independent and identically distributed according to some multivariate sub-Gaussian distribution F . Then there exists a sequence of $d \times d$ orthogonal matrices \widetilde{W}_n , such that for any fixed index i with $Z_i = z_i$, we have*

$$n^{1/2} \widetilde{W}_n^\top [(\widehat{B} - B) U_B S_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \Sigma(z_i))$$

where $\Sigma(z_i) = \Xi^{-1} \widetilde{\Sigma}(z_i) \Xi^{-1}$, $\Xi = \mathbb{E}[Z_k Z_k^\top] \in \mathbb{R}^{d \times d}$, $\mu = \mathbb{E}[Z_k] \in \mathbb{R}^d$ and

$$\widetilde{\Sigma}(z_i) = \mathbb{E}_{Z_k} [(\sigma^2 \|z_i - Z_k\|^2 + \mathbb{E}[E_{ij}^3] \|z_i - Z_k\| + \frac{1}{4} \mathbb{E}[E_{ij}^4] - \frac{\sigma^4}{4})(Z_k - \mu)(Z_k - \mu)^\top] \in \mathbb{R}^{d \times d}$$

is a covariance matrix depending on z . Here $(A)_i$ or $[A]_i$ denote the i th row of a matrix A .

Proof: Recall that $PZ = U_B S_B^{1/2} \widetilde{W}_n$. We therefore have

$$\begin{aligned} n^{1/2} \widetilde{W}_n^\top [(\widehat{B} - B) U_B S_B^{-1/2}]_i &= n^{1/2} \widetilde{W}_n^\top [(\widehat{B} - B) PZ \widetilde{W}_n^\top S_B^{-1}]_i \\ &= n^{1/2} \widetilde{W}_n^\top S_B^{-1} \widetilde{W}_n [(\widehat{B} - B) PZ]_i \\ &= -n^{1/2} \widetilde{W}_n^\top S_B^{-1} \widetilde{W}_n [P(D \circ E + \frac{E^2}{2}) PZ]_i \\ &= -n^{1/2} \widetilde{W}_n^\top S_B^{-1} \widetilde{W}_n \left[P \left(D \circ E + \frac{E^2 - \sigma^2 1_n 1_n^\top}{2} \right) PZ \right]_i \end{aligned}$$

CHAPTER 2. CMDS WITH PERTURBATION

The last equality holds since $P1_n = 0$. Now

$$PZ = Z - 1_n \bar{Z} = Z - 1_n \mu^\top + \tilde{R}_n$$

where $\|\tilde{R}_n\| = O(n^{-1/2})$ with high probability. Therefore,

$$\begin{aligned} n^{1/2} \tilde{W}_n^\top [(\hat{B} - B)U_B S_B^{-1/2}]_i &= -n \tilde{W}_n^\top S_B^{-1} \tilde{W}_n \left[n^{-1/2} \sum_{j \neq i}^n \left(D_{ij} E_{ij} + \frac{E_{ij}^2 - \sigma^2 1_n 1_n^\top}{2} \right) (Z_j - \mu) \right] \\ &\quad + o(1) \end{aligned}$$

Conditioning on $Z_i = z_i$ and ignoring the term $o(1)$ that vanishes as $n \rightarrow \infty$, the above expression is sum of $n - 1$ independent mean 0 random vector. We then invoke the Lindeberg-Feller central limit theorem to show that this sum converges to a multivariate normal. We now evaluate the covariance matrix for this sum. Each summand has covariance matrix of the form

$$\text{cov} \left[\left(E_{ij} D_{ij} + \frac{E_{ij}^2 - \sigma^2}{2} \right) (Z_j - \mu) \right] = \text{Var} \left(E_{ij} \|z_i - Z_j\| + \frac{E_{ij}^2 - \sigma^2}{2} \right) (Z_j - \mu)(Z_j - \mu)^\top.$$

Since $\mathbb{E}[E_{ij}] = 0$ and $\mathbb{E}[E_{ij}^2] = \sigma^2$, we also have

$$\text{Var} \left(E_{ij} \|z_i - Z_j\| + (E_{ij}^2 - \sigma^2)/2 \right) = \mathbb{E} \left[E_{ij}^2 \|z_i - Z_j\|^2 + E_{ij} \|z_i - Z_j\| (E_{ij}^2 - \sigma^2) + \frac{(E_{ij}^2 - \sigma^2)^2}{4} \right]$$

where the expectation is taken with respect to E_{ij} and conditional on Z_j . Averaging

CHAPTER 2. CMDS WITH PERTURBATION

over the indices j and then taking the limit as $n \rightarrow \infty$ yields

$$\begin{aligned}\tilde{\Sigma}_n(z_i) &= \text{Var} \left[n^{-1/2} \sum_{j \neq i}^n \left(D_{ij} E_{ij} + \frac{E_{ij}^2 - \sigma^2 \mathbf{1}_n \mathbf{1}_n^\top}{2} \right) (Z_j - \mu) \right] \\ &\longrightarrow \mathbb{E}_{Z_k} \left[\left(\sigma^2 \|z_i - X_k\|^2 + \mathbb{E}[E_{ij}^3] \|z_i - Z_k\| + \frac{1}{4} \mathbb{E}[E_{ij}^4] - \frac{\sigma^4}{4} \right) (Z_k - \mu)(Z_k - \mu)^\top \right].\end{aligned}$$

By the strong law of large numbers, we have

$$\frac{\widetilde{W}_n^\top S_B \widetilde{W}_n}{n} = \frac{1}{n} Z^\top P Z \rightarrow \Xi \in \mathbb{R}^{d \times d}$$

almost surely. Hence $(n \widetilde{W}_n^\top S_B^{-1} \widetilde{W}_n) \rightarrow \Xi^{-1}$ almost surely. Slutsky's theorem implies

$$n^{1/2} \widetilde{W}_n^\top [(\widehat{B} - B) U_B S_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \Xi^{-1} \widetilde{\Sigma}(z_i) \Xi^{-1})$$

as desired. ■

Finally we state the following lemma showing that any row of these matrices, when scaled by $n^{1/2}$, converges to 0 in probability.

CHAPTER 2. CMDS WITH PERTURBATION

Lemma 2.6.4 *For any fixed index i , we have, simultaneously*

$$n^{1/2}[(\widehat{B} - B)U_B(W^*S_{\widehat{B}}^{-1/2} - S_B^{-1/2}W^*)]_i \xrightarrow{P} 0 \quad (2.6)$$

$$n^{1/2}[U_B U_B^\top (\widehat{B} - B)U_B W^* S_{\widehat{B}}^{-1/2}]_i \xrightarrow{P} 0 \quad (2.7)$$

$$n^{1/2}[(I - U_B U_B^\top)(\widehat{B} - B)(\widehat{U}_B - U_B W^*)S_{\widehat{B}}^{-1/2}]_i \xrightarrow{P} 0 \quad (2.8)$$

$$n^{1/2}[U_B(U_B^\top U_{\widehat{B}} - W^*)S_{\widehat{B}}^{1/2}]_i \xrightarrow{P} 0. \quad (2.9)$$

$$n^{1/2}[U_B(W^*S_{\widehat{B}}^{1/2} - S_B^{1/2}W^*)]_i \xrightarrow{P} 0. \quad (2.10)$$

The rest of this section is devoted toward proving Lemma 2.6.4, for which we need the following technical lemmas controlling the spectral norm of $\|\widehat{B} - B\|$ and $\|U_B^\top \widehat{U}_B - W^*\|$ (recall that W^* is the closest orthogonal matrix, in Frobenius norm, to $U_B^\top \widehat{U}_B$.) We start with a bound for the spectral norm of $B - \widehat{B}$.

Proposition 2.6.5 $\|B - \widehat{B}\| = \mathcal{O}(\sqrt{n \log n})$ with high probability.

Proof: We have

$$\begin{aligned} \|B - \widehat{B}\| &= \left\| -\frac{1}{2}PD^2P + \frac{1}{2}P(D + E)^2P \right\| \\ &= \|PD \circ EP + \frac{1}{2}PE^2P\| \text{ (where } \circ \text{ is the Hadamard product)} \\ &\leq \|D \circ E\| + \frac{1}{2}\|E^2 - \mathbb{E}[E^2]\| \text{ (since } \|P\| = 1.) \\ &= \mathcal{O}(\sqrt{n}) + \mathcal{O}(\sqrt{n \log n}) \end{aligned}$$

CHAPTER 2. CMDS WITH PERTURBATION

Note that here we used $\mathbb{E}[D \circ E] = 0$ and $\mathbb{E}[\frac{1}{2}PE^2P] = 0$. Each entries of $D \circ E$ is of sub-Gaussian distribution with mean 0 and each entries of $E^2 - \mathbb{E}[E^2]$ is of sub-exponential distribution with mean 0. An application of Theorem 4.4.5 in Vershynin [2018] and Matrix Bernstein for the sub-exponential case gives the desired result. ■

Lemma 2.6.6 *Let $X_1, \dots, X_n, Y \stackrel{i.i.d}{\sim} F$ for some sub-Gaussian distribution F , where X_i is the i th row of the configuration matrix X of B viewed as a column vector. Let $\Xi = \mathbb{E}[X_1 X_1^\top]$ be of rank d , then $\lambda_i(B) = \Omega(n)$ almost surely.*

Proof: For any matrix H , the nonzero eigenvalues of $H^\top H$ are the same as those HH^\top , so $\lambda_i(XX^\top) = \lambda_i(X^\top X)$. In what follows, we remind the reader that X is a matrix whose rows are the transposes of the column vectors X_i , and Y is a d -dimensional vector that is independent from and has the same distribution as that of the X_i . We observe that

$$(X^\top X - n\mathbb{E}[YY^\top])_{ij} = \sum_{k=1}^n (X_{ki}X_{kj} - \mathbb{E}[Y_i Y_j])$$

is a sum of n independent mean-zero sub-Gaussian random variables. By a general Hoeffding's inequality for sub-gaussian random variables [Vershynin, 2018], for all $i, j \in [d]$,

$$\mathbb{P}[|(X^\top X - n\mathbb{E}[YY^\top])_{ij}| \geq t] \leq 2 \exp\left\{\frac{-ct^2}{nM}\right\},$$

CHAPTER 2. CMDS WITH PERTURBATION

where $M = \max_k \|(X_{ki}X_{kj} - \mathbb{E}[Y_iY_j])\|_{\varphi_2}^2$. Therefore,

$$\mathbb{P}[|(X^\top X - n\mathbb{E}[YY^\top])_{ij}| \geq C\sqrt{n \log n}] \leq 2n^{-\frac{2C^2}{M^2}}.$$

A union bound over all $i, j \in [d]$ implies that $\|X^\top X - n\mathbb{E}[YY^\top]\|_F^2 \leq C^2 d^2 n \log n$ with probability at least $1 - 2n^{-2C^2/M^2}$, i.e. $\|X^\top X - n\mathbb{E}[YY^\top]\|_F \leq Cd\sqrt{n \log n}$ with high probability for any $C > \frac{M}{\sqrt{2}}$. By the Hoffman-Wielandt inequality, $|\lambda_i(XX^\top) - n\lambda_i(\mathbb{E}[YY^\top])| \leq Cd\sqrt{n \log n}$, and by reverse triangle inequality, we obtain

$$\lambda_i(XX^\top) \geq \lambda_d(XX^\top) \geq |n\lambda_d(\Xi)| - Cd\sqrt{n \log n} = \Omega(n)$$

holds almost surely. ■

Proposition 2.6.7 *Let $W_1 \Sigma W_2^\top$ be the singular value decomposition of $U_B^\top U_{\hat{B}}$, then with high probability, $\|U_B^\top U_{\hat{B}} - W_1 W_2^\top\| = \mathcal{O}(n^{-1} \log n)$.*

Proof: Let $\sigma_1, \sigma_2, \dots, \sigma_d$ be the singular values of $U_B^\top U_{\hat{B}}$ (the diagonal entries of Σ).

Then $\sigma_i = \cos(\theta_i)$ where θ_i 's are the principal angles between the subspace spanned by U_B and $U_{\hat{B}}$. The Davis-Kahan $\sin(\Theta)$ theorem [Davis and Kahan, 1970] gives

$$\|U_{\hat{B}} U_{\hat{B}}^\top - U_B U_B^\top\| = \max_i |\sin(\theta_i)| \leq \frac{C\|B - \hat{B}\|}{\lambda_d(B)} = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right)$$

for sufficiently large n . Note in the last equality we used the previous two lemmas.

CHAPTER 2. CMDS WITH PERTURBATION

Thus,

$$\begin{aligned} \|U_B^\top U_{\hat{B}} - W_1 W_2^\top\|_F &= \|\Sigma - I\|_F = \sqrt{\sum_{i=1}^d (1 - \sigma_i)^2} \leq \sum_{i=1}^d (1 - \sigma_i) \leq \sum_{i=1}^d (1 - \sigma_i^2) \\ &= \sum_{i=1}^d \sin(\theta_i)^2 \leq d \|U_{\hat{B}} U_{\hat{B}}^\top - U_B U_B^\top\|^2 = \mathcal{O}\left(\frac{\log n}{n}\right) \end{aligned}$$

■

Recall that a random vector X is sub-exponential if $\mathbb{P}[|X| > t] \leq 2e^{-\frac{t}{K}}$ for some constant K and for all $t \geq 0$. Associated with a sub-exponential random variable there is a Orlicz norm defined as $\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E} \exp(\frac{|X|}{t}) \leq 2\}$. Furthermore, a random variable X is sub-Gaussian if and only if X^2 is sub-exponential, and $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$. We now have the following lemma which allows us to juxtapose the ordering in the matrix product $W^* \hat{S}_B$ and $S_B W^*$ (and similarly $W^* \hat{S}_B^{1/2}$ and $S_B^{1/2} W^*$.) This juxtaposition is essential in showing Eq. (2.6) and Eq. (2.10) in Lemma 2.6.4.

Lemma 2.6.8 *Let $W^* = W_1 W_2^\top$. Then with high probability,*

$$\|W^* S_{\hat{B}} - S_B W^*\|_F = \mathcal{O}(\log n); \quad \text{and} \quad \|W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*\|_F = \mathcal{O}(n^{-\frac{1}{2}} \log n).$$

Proof: Let $R = U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}$. Note R is the residual after projecting $U_{\hat{B}}$ orthogonally onto the column space of U_B , and thus $\|U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}\|_F \leq \min_W \|U_{\hat{B}} - U_B W\|_F$ where the minimization is over all orthogonal matrices W . By a variant of the Davis-

CHAPTER 2. CMDS WITH PERTURBATION

Kahan $\sin \Theta$ theorem [Yu et al., 2015], we have

$$\min_W \|U_B W - U_{\hat{B}}\|_F \leq \frac{C\sqrt{d}\|B - \hat{B}\|}{\lambda_d(B)},$$

and hence $\|R\|_F \leq \mathcal{O}(\sqrt{\frac{\log n}{n}})$. Now consider

$$\begin{aligned} W^* S_{\hat{B}} &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top U_{\hat{B}} S_{\hat{B}} \\ &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top \hat{B} U_{\hat{B}} \\ &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top (\hat{B} - B) U_{\hat{B}} + U_B^\top B U_{\hat{B}} \\ &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top (\hat{B} - B) R + U_B^\top (\hat{B} - B) U_B U_B^\top U_{\hat{B}} + S_B U_B^\top U_{\hat{B}}. \end{aligned}$$

Note here we use the fact $U_{\hat{B}} S_{\hat{B}} = \hat{B} U_{\hat{B}}$. Now write

$$S_B U_B^\top U_{\hat{B}} = S_B (U_B^\top U_{\hat{B}} - W^*) + S_B W^*,$$

then we have

$$W^* S_{\hat{B}} - S_B W^* = (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top (\hat{B} - B) R + U_B^\top (\hat{B} - B) U_B U_B^\top U_{\hat{B}} + S_B (U_B^\top U_{\hat{B}} - W^*).$$

CHAPTER 2. CMDS WITH PERTURBATION

This gives

$$\begin{aligned}
\|W^*S_{\hat{B}} - S_BW^*\|_F &\leq \|(U_B^\top U_{\hat{B}} - W^*)(S_{\hat{B}} + S_B)\|_F + \|U_B^\top(\hat{B} - B)R\|_F \\
&\quad + \|U_B^\top(\hat{B} - B)U_B U_B^\top U_{\hat{B}}\|_F \\
&\leq \|(U_B^\top U_{\hat{B}} - W^*)\|_F(\|S_{\hat{B}}\| + \|S_B\|) + \|U_B^\top(\hat{B} - B)R\|_F \\
&\quad + \|U_B^\top(\hat{B} - B)U_B U_B^\top U_{\hat{B}}\|_F \\
&\leq \|W_1W_2^\top - U_B^\top U_{\hat{B}}\|_F(\mathcal{O}(n) + \mathcal{O}(n)) + \|U_B^\top(\hat{B} - B)R\|_F \\
&\quad + \|U_B^\top(\hat{B} - B)U_B\|_F \\
&\leq \mathcal{O}(n^{-1})(\mathcal{O}(n) + \mathcal{O}(n)) + \mathcal{O}(\log n) + \|U_B^\top(\hat{B} - B)U_B\|_F \\
&= \mathcal{O}(\log n) + \|U_B^\top(\hat{B} - B)U_B\|_F.
\end{aligned}$$

Now consider the term $U_B^\top(\hat{B} - B)U_B \in \mathbb{R}^{d \times d}$. If we denote U_i be the i th column of U_B , then for each i, j th entry, we have

$$(U_B^\top(\hat{B} - B)U_B)_{ij} = U_i^\top(\hat{B} - B)U_j = \frac{1}{2}V_i^\top(\Delta^2 - D^2)V_j$$

where $V = PU_B$. Furthermore, we have

$$V_i^\top(\Delta^2 - D^2)V_j = \sum_{k,l} V_{ik}(\Delta_{kl}^2 - D_{kl}^2)V_{jl}. \quad (2.11)$$

Recall, since X_k 's are sub-Gaussian, thus equation (2.11) is a sum of mean zero sub-

CHAPTER 2. CMDS WITH PERTURBATION

exponential random variables. By Bernstein's inequality [Vershynin, 2018], we have

$$\mathbb{P}\left[\left|\sum_{k,l}(\Delta_{kl}^2 - D_{kl}^2)V_{ik}V_{jl}\right| > t\right] \leq 2 \exp\left\{-C \min\left(\frac{t^2}{M^2 \sum_{k,l} V_{ik}^2 V_{jl}^2}, \frac{t}{M \max_{k,l}(V_{ik}V_{jl})}\right)\right\}$$

where $M := \max_{k,l} \|\Delta_{kl}^2 - D_{kl}^2\|_{\psi_1}$. Since $\sum_k V_{ik}^2 \leq 1 \forall i$, we have that each entry of $U_B^\top(\hat{B} - B)U_B \in \mathbb{R}^{d \times d}$ is $\mathcal{O}(\log n)$, and

$$\|U_B^\top(\hat{B} - B)U_B\|_F = \mathcal{O}(\log n). \quad (2.12)$$

This then gives $\|W^*S_{\hat{B}} - S_B W^*\|_F = \mathcal{O}(\log n)$, with high probability.

Finally, consider $\|W^*S_{\hat{B}}^{1/2} - S_B^{1/2}W^*\|_F$. The i, j th entry of $W^*S_{\hat{B}}^{1/2} - S_B^{1/2}W^*$ is

$$\begin{aligned} W^*_{ij}(\lambda_j^{1/2}(\hat{B}) - \lambda_j^{1/2}(B)) &= W^*_{ij} \frac{\lambda_j(\hat{B}) - \lambda_j(B)}{\lambda_j^{1/2}(\hat{B}) + \lambda_j^{1/2}(B)} \leq W^*_{ij} \frac{\lambda_j(\hat{B}) - \lambda_j(B)}{\Omega(\sqrt{n})} \\ &= \mathcal{O}(n^{-\frac{1}{2}} \log n), \end{aligned}$$

as desired (note in the last inequality, we used the first part of this Lemma. \blacksquare We

now proceed to prove Lemma 2.6.4.

Proof: [Proof of Lemma 2.6.4]

CHAPTER 2. CMDS WITH PERTURBATION

To show Eq. (2.6), we have

$$\begin{aligned}
\sqrt{n}\|(\hat{B} - B)U_B(W^*S_{\hat{B}}^{-1/2} - S_B^{-1/2}W^*)\|_F &\leq \sqrt{n}\|(\hat{B} - B)U_B\| \times \|W^*S_{\hat{B}}^{-1/2} - S_B^{-1/2}W^*\|_F \\
&\leq \sqrt{n}\|(\hat{B} - B)\| \times \|W^*S_{\hat{B}}^{-1/2} - S_B^{-1/2}W^*\|_F \\
&= \sqrt{n}\mathcal{O}(\sqrt{n \log n})\mathcal{O}(n^{-\frac{3}{2}} \log n) = \frac{C \log n \sqrt{\log n}}{\sqrt{n}}
\end{aligned}$$

which converges to 0 as $n \rightarrow \infty$.

Let us now consider Eq. (2.7). Recall that $X = U_B S_B^{1/2} W$ for some orthogonal matrix W , and since X_i 's are sub-Gaussian, $\|X_i\|$ is bounded by some constant C with high probability, i.e., $\|X_i\| = \sqrt{\sum_{j=1}^d \sigma_j U_{Bij}^2} \leq C$ with high probability, where σ_i 's are the diagonal entries of $S_B^{1/2}$. Note that $\sigma_i = \Omega(n) \geq C'n$ for all i and some constant C' . We thus obtain $\sqrt{\sum_{j=1}^d U_{Bij}^2} \leq \frac{C}{\sqrt{n}}$, i.e., $\|U_B\|_{2 \rightarrow \infty} \leq \frac{C}{\sqrt{n}}$. Hence,

$$\begin{aligned}
\|[U_B U_B^\top (\hat{B} - B) U_B W^* S_{\hat{B}}^{-1/2}]_h\| &\leq \|U_B\|_{2 \rightarrow \infty} \|U_B^\top (\hat{B} - B) U_B\| \times \|S_{\hat{B}}^{-1/2}\| \\
&\leq \frac{C}{\sqrt{n}} \mathcal{O}(\log n) \mathcal{O}(n^{-\frac{1}{2}}) \leq \frac{C \log n}{n}
\end{aligned}$$

which also converges to 0 as $n \rightarrow \infty$ (note in the last inequality we used 2.12).

To show Eq. (2.8), we must bound $\|[(I - U_B U_B^\top)(\hat{B} - B)(\hat{U}_B - U_B W^*)S_{\hat{B}}^{-1/2}]_h\|$.

CHAPTER 2. CMDS WITH PERTURBATION

Define

$$\begin{aligned} G_1 &= (I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}S_{\widehat{B}}^{-1/2}, \\ G_2 &= (I - U_B U_B^\top)(\widehat{B} - B)U_B(U_B^\top U_{\widehat{B}} - W^*)S_{\widehat{B}}^{-1/2} \end{aligned}$$

Note that $(I - U_B U_B^\top)(\widehat{B} - B)(\widehat{U}_B - U_B W^*)S_{\widehat{B}}^{-1/2} = G_1 + G_2$. We now only need to bound the h th row of G_1 and G_2 .

$$\begin{aligned} \|G_2\|_F &\leq \|(I - U_B U_B^\top)(\widehat{B} - B)U_B\| \times \|U_B^\top U_{\widehat{B}} - W^*\|_F \times \|S_{\widehat{B}}^{-\frac{1}{2}}\| \\ &\leq \|(I - U_B U_B^\top)\| \times \|\widehat{B} - B\| \times \|U_B^\top U_{\widehat{B}} - W^*\|_F \times \|S_{\widehat{B}}^{-\frac{1}{2}}\| \\ &= \mathcal{O}(1)\mathcal{O}(\sqrt{n \log n})\mathcal{O}(n^{-1})\mathcal{O}(n^{-\frac{1}{2}}) = \mathcal{O}\left(\frac{\sqrt{\log n}}{n}\right) \end{aligned}$$

Thus $\|\sqrt{n}G_2\|_F$ converges to 0 as $n \rightarrow \infty$. We now consider the rows of G_1 . Note that $U_{\widehat{B}}^\top U_{\widehat{B}} = I$ and hence

$$\begin{aligned} \|(G_1)_h\| &= \|(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}S_{\widehat{B}}^{-1/2}\|_h \\ &= \|(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}U_{\widehat{B}}^\top U_{\widehat{B}}S_{\widehat{B}}^{-1/2}\|_h \\ &= \|U_{\widehat{B}}S_{\widehat{B}}^{-1/2}\| \times \|(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}U_{\widehat{B}}^\top\|_h \\ &\leq \frac{C}{\sqrt{n}} \|(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}U_{\widehat{B}}^\top\|_h \end{aligned}$$

CHAPTER 2. CMDS WITH PERTURBATION

Let $\epsilon > 0$ be a constant, we now want to show

$$\lim_{n \rightarrow \infty} \mathbb{P}[\|\sqrt{n}(G_1)_h\| > \epsilon] = 0. \quad (2.13)$$

To see this, define

$$H_1 = (I - U_B U_B^\top)(\hat{B} - B)(I - U_B U_B^\top)U_{\hat{B}}U_{\hat{B}}^\top.$$

We will show that

$$\lim_{n \rightarrow \infty} \mathbb{P}[\|\sqrt{n}(H_1)_h\| > n^{1/4}] = 0 \quad (2.14)$$

which will imply equation 2.13.

We will show 2.14, by verifying that the Frobenius norms of the rows of H_1 are exchangeable, and hence all of these Frobenius norms have the same expectation, then Markov's inequality will bound the probability that the Frobenius norm of any fixed row exceeds a specified threshold. To prove the exchangeability, note that if $Q \in \mathbb{R}^{n \times n}$ is any permutation matrix, right multiplication of an $m \times m$ matrix J by Q^\top only permutes the columns of J , hence the Frobenius norm of the i th row of QJQ^\top is the same as the Frobenius norm of the i th row of QG . For any $n \times n$ real symmetric matrix J , let $\mathcal{P}_d(J)$ denote the projection onto the eigenspace defined by the top d eigenvalues (in magnitude) of J . Similarly, let $\mathcal{P}_d^\perp(J)$ denote the projection onto the orthogonal complement of that eigenspace. For the matrix B , the columns of

CHAPTER 2. CMDS WITH PERTURBATION

U_B are a basis for the eigenspace associated to the top d eigenvalues, and the matrix $U_B U_B^\top$ is the unique projection operator into this eigenspace.

Observe that

$$QU_B U_B^\top Q^\top Q B Q^\top = QU_B U_B^\top B Q^\top,$$

and thus $QU_B U_B^\top Q^\top$ is the projection matrix onto the eigenspace of top d eigenvalues of $Q B Q^\top$ if and only if $U_B U_B^\top$ is the corresponding projection matrix for B . Similarly, $U_{\hat{B}} U_{\hat{B}}^\top$ is the unique projection operator onto the eigenspace defined by the top d eigenvalues (in magnitude) of \hat{B} , and $QU_{\hat{B}} U_{\hat{B}}^\top Q^\top$ is the corresponding projection matrix for $Q \hat{B} Q^\top$.

Now, for any pair of $n \times n$ matrices (J, K) , let $\mathcal{L}(J, K)$ represent the following operator:

$$\mathcal{L}(J, K) = \mathcal{P}_d^\perp(J)(J - K)\mathcal{P}_d^\perp(J)\mathcal{P}_d(K)$$

It is easy to see that $\mathcal{L}(\hat{B}, B) = H_1$, and by uniqueness of projection, we note that

$$\mathcal{L}(Q \hat{B} Q^\top, Q B Q^\top) \tag{2.15}$$

$$= Q(I - U_B U_B^\top)Q^\top Q(\hat{B} - B)Q^\top Q(I - U_B U_B^\top)Q^\top Q(U_{\hat{B}} U_{\hat{B}}^\top)Q^\top \tag{2.16}$$

$$= Q H_1 Q^\top \tag{2.17}$$

Since the Z_i are i.i.d., the entries of the matrix pair (\hat{B}, B) have the same joint distribution as the entries of the pair $(Q \hat{B} Q^\top, Q B Q^\top)$, hence the entries of the matrix

CHAPTER 2. CMDS WITH PERTURBATION

$\mathcal{L}(\widehat{B}, B)$ have the same distribution as those of $\mathcal{L}(Q\widehat{B}Q^\top, QBQ^\top)$. By equation 2.15, this implies that H_1 has the same distribution as QH_1Q^\top . Since the Frobenius norm of any row of QH_1Q^\top is exactly equal to the Frobenius norm of the corresponding row of QH_1 , we conclude that the Frobenius norm of rows of H_1 have the same distribution as the Frobenius norms of the rows of QH_1 , establishing that the Frobenius norms of the rows of H_1 are exchangeable. This row-exchangeability for the Frobenius norms of H_1 implies that each row has the same expectation, and hence for any fixed index h , $n\mathbb{E}\|(H_1)_h\|^2 = \mathbb{E}\|H_1\|_F^2$.

Markov's inequality then implies

$$\begin{aligned} \mathbb{P}[\|\sqrt{n}(H_1)_h\| > t] &\leq \frac{n\mathbb{E}\|[(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}U_{\widehat{B}}^\top]_h\|^2}{t^2} \\ &= \frac{\mathbb{E}(\|(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}U_{\widehat{B}}^\top\|_F^2)}{t^2} \end{aligned}$$

Furthermore,

$$\|(I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}}U_{\widehat{B}}^\top\|_F \leq \|\widehat{B} - B\| \times \|U_{\widehat{B}} - U_B U_B^\top U_{\widehat{B}}\|_F$$

We now recall the following two observations

- The optimization problem $\min_{T \in \mathbb{R}^{d \times d}} \|U_{\widehat{B}} - U_B T\|_F^2$ is solved by $T = U_B^\top U_{\widehat{B}}$.
- By theorem 2 of Yu et al. [2015], there exists $W \in \mathbb{R}^{d \times d}$ orthogonal, such that

$$\|U_{\widehat{B}} - U_B W\|_F \leq C \|U_{\widehat{B}} U_{\widehat{B}}^\top - U_B U_B^\top\|_F.$$

CHAPTER 2. CMDS WITH PERTURBATION

Combining the two facts above, we conclude that $\|U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}\|_F^2 \leq \frac{C}{n}$ with high probability, as in Lemma 2.6.8, hence

$$\|(I - U_B U_B^\top)(\hat{B} - B)(I - U_B U_B^\top)U_{\hat{B}}U_{\hat{B}}^\top\|_F \leq \mathcal{O}(\sqrt{n \log n}) \frac{C}{\sqrt{n}} = \mathcal{O}(\sqrt{\log n}),$$

with high probability. Therefore,

$$\mathbb{P}(\|\sqrt{n}(H_1)_h\| > t) \leq \frac{\sqrt{\log n}}{t^2}.$$

picking $t = n^{\frac{1}{4}}$, we get $\lim_{n \rightarrow \infty} Cn^{-1/2}\|\sqrt{n}(H_1)_h\| = 0$.

Finally, Eq. (2.9) and Eq. (2.10) follow from Lemma 2.6.7 and Lemma 2.6.8 and the bound $\|U_B\|_{2 \rightarrow \infty} \leq Cn^{-1/2}$. ■

2.6.2 Adaptation for Theorem 2.4.1 and 2.4.4

The major difference between our main theorems is the calculation of the covariance matrices. In this section, we will give those calculations.

Lemma 2.6.9 *Let Z_1, \dots, Z_n be independent and identically distributed according to some multivariate sub-Gaussian distribution F and let our model be as in Theorem 2.4.1. Then there exists a sequence of $d \times d$ orthogonal matrices \widetilde{W}_n , such that for*

CHAPTER 2. CMDS WITH PERTURBATION

any fixed index i , we have

$$n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)U_B S_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \Sigma)$$

where $\Sigma = \frac{\sigma^2}{4}\Xi^{-1}$, $\Xi = \text{cov}(Z_k)$. Here $(A)_i$ or $[A]_i$ denote the i th row of a matrix A .

Proof: Recall that $PZ = U_B S_B^{1/2}\widetilde{W}_n$. We therefore have

$$\begin{aligned} & n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)U_B S_B^{-1/2}]_i \\ &= n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)PZ\widetilde{W}_n^\top S_B^{-1}]_i \\ &= n^{1/2}\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n[(\widehat{B} - B)PZ]_i \\ &= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n[P(D^{(2)} - \Delta^{(2)})PZ]_i \\ &= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n\left[P(D^{(2)} - \Delta^{(2)})(I - 1_n 1_n^\top/n)Z\right]_i \\ &= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n\left[P(D^{(2)} - \Delta^{(2)})(Z - 1_n \mu^\top)\right]_i \\ &\quad (\text{ since } PZ = Z - 1_n \bar{Z} = Z - 1_n \mu^\top) \\ &= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n\left[(D^{(2)} - \Delta^{(2)})(Z - 1_n \mu^\top) \right. \\ &\quad \left. - 1_n 1_n^\top/n(D^{(2)} - \Delta^{(2)})(Z - 1_n \mu^\top)\right]_i \\ &\quad (\text{note that } \frac{1_n 1_n^\top}{n}(D^{(2)} - \Delta^{(2)})(Z - 1_n \mu^\top) \rightarrow 0 \text{ as } n \rightarrow \infty) \\ &= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n\left[(D^{(2)} - \Delta^{(2)})(Z - 1_n \mu^\top)\right]_i \text{ as } n \rightarrow \infty \end{aligned}$$

CHAPTER 2. CMDS WITH PERTURBATION

Therefore,

$$n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)U_BS_B^{-1/2}]_i = \frac{1}{2}n\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n \left[n^{-1/2} \sum_{j \neq i}^n \left(D_{ij}^{(2)} - \Delta_{ij}^{(2)} \right) (Z_j - \mu) \right]$$

Conditioning on $Z_i = z_i$, the above expression is sum of $n - 1$ independent mean 0 random vectors. We then invoke the Lindeberg-Feller central limit theorem to show that this sum converges to a multivariate normal. We now evaluate the covariance matrix for this sum. Each summand has covariance matrix of the form

$$\begin{aligned} \text{cov} \left[(D_{ij}^{(2)} - \Delta_{ij}^{(2)})(Z_j - \mu) \right] &= \text{Var} \left(D_{ij}^{(2)} - \Delta_{ij}^{(2)} \right) (Z_j - \mu)(Z_j - \mu)^\top \\ &= \sigma^2 (Z_j - \mu)(Z_j - \mu)^\top \end{aligned}$$

$$\text{Since } \mathbb{E}[E_{ij}] = 0 \text{ and } \mathbb{E}[E_{ij}^2] = \sigma^2$$

. By the strong law of large numbers, we have

$$\frac{\widetilde{W}_n^\top S_B \widetilde{W}_n}{n} = \frac{1}{n} Z^\top P Z \rightarrow \Xi \in \mathbb{R}^{d \times d}$$

almost surely. Hence $(n\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n) \rightarrow \Xi^{-1}$ almost surely. Slutsky's theorem implies

$$n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)U_BS_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \frac{\sigma^2}{4}\Xi^{-1})$$

as desired. ■

CHAPTER 2. CMDS WITH PERTURBATION

Lemma 2.6.10 *Let Z_1, \dots, Z_n be independent and identically distributed according to some multivariate sub-Gaussian distribution F and let our model be as in Theorem 2.4.4. Then there exists a sequence of $d \times d$ orthogonal matrices \widetilde{W}_n , such that for any fixed index i , we have*

$$n^{1/2} \widetilde{W}_n^\top [(\widehat{B} - q_n B) \frac{1}{\sqrt{q_n}} U_B S_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \Sigma)$$

Here $\Sigma(z) = \Xi^{-1} \widetilde{\Sigma}(z) \Xi^{-1}$, $\Xi = \text{cov}(Z_i) \in \mathbb{R}^{d \times d}$ and with $\mu = \mathbb{E}[Z_i] \in \mathbb{R}^d$,

$$\widetilde{\Sigma}(z) = \frac{1 - q_n}{4} \times \mathbb{E}_{Z_k} \left[\|z - Z_k\|^4 (Z_k - \mu)(Z_k - \mu)^\top \right]$$

is a covariance matrix depending on z , and $(A)_i$ or $[A]_i$ denote the i th row of a matrix A .

Proof: First recall that as our mode we have

$$E_{ij} = \begin{cases} -D_{ij}, & \text{with probability } 1 - q_n \\ 0, & \text{with probability } q_n \end{cases}$$

hence it is easy to see that $\mathbb{E}[E] = D(q_n - 1)$, $\mathbb{E}[E^2] = D^2(1 - q_n)$ and furthermore,

$$\mathbb{E}[\widehat{B}] = q_n B.$$

CHAPTER 2. CMDS WITH PERTURBATION

As in Lemma 2.6.9, we will investigate the term

$$\begin{aligned}
& n^{1/2} \widetilde{W}_n^\top \left[(\widehat{B} - q_n B) \frac{1}{\sqrt{q_n}} U_B S_B^{-1/2} \right]_i \\
&= n^{1/2} \widetilde{W}_n^\top \left[(\widehat{B} - q_n B) \frac{1}{\sqrt{q_n}} (PZ) \widetilde{W}_n^\top S_B^{-1/2} \right]_i \\
&= n^{1/2} \widetilde{W}_n^\top S_B^{-1/2} \widetilde{W}_n \left[(\widehat{B} - q_n B) \frac{1}{\sqrt{q_n}} (PZ) \right]_i \\
&= -\sqrt{n} \widetilde{W}_n^\top S_B^{-1/2} \widetilde{W}_n \left[\frac{1}{\sqrt{q_n}} (D \circ E - D^2(q_n - 1) + \frac{E^2 - D^2(1 - q_n)}{2})(I - \mathbf{1}_n \mathbf{1}_n^\top / n) Z \right]_i \\
&= -n \widetilde{W}_n^\top S_B^{-1/2} \widetilde{W}_n \left[\frac{1}{\sqrt{q_n}} \frac{1}{\sqrt{n}} \left(\sum_{i \neq j}^n (D \circ E - D^2(q_n - 1) + \frac{E^2 - D^2(1 - q_n)}{2})_{ij} (Z_j - \mu) \right) \right. \\
&\quad \left. - \frac{1}{\sqrt{q_n}} \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n D \circ E - D^2(q_n - 1) + \frac{E^2 - D^2(1 - q_n)}{2} \right)_{ii} (Z_j - \mu) \right] \\
&= -n \widetilde{W}_n^\top S_B^{-1/2} \widetilde{W}_n \left[\frac{1}{\sqrt{q_n}} \frac{1}{\sqrt{n}} \left(\sum_{i \neq j}^n (D \circ E - D^2(q_n - 1) + \frac{E^2 - D^2(1 - q_n)}{2})_{ij} (Z_j - \mu) \right) \right]
\end{aligned}$$

Note that the second half of the above expression is 0, hence leave us the above expression. Furthermore, the above term is a sum of $n - 1$ independent mean 0 random vectors, which we will now calculate its Cov matrix.

Condition on $Z_i = z_i$, we have:

$$\begin{aligned}
& \text{Cov} \left[\frac{1}{\sqrt{q_n}} \left(D_{ij} E_{ij} - D_{ij}^2(q_n - 1) + \frac{E_{ij}^2 - D_{ij}^2(1 - q_n)}{2} \right)_{ij} (Z_j - \mu) \right] \\
&= \frac{1}{q_n} \text{Var} \left(D_{ij} E_{ij} - D_{ij}^2(q_n - 1) + \frac{E_{ij}^2 - D_{ij}^2(1 - q_n)}{2} \right) (z_j - \mu)(Z_j - \mu)^\top \\
&= \frac{1}{q_n} \text{Var} \left(\|z_i - Z_j\| E_{ij} - \|z_i - Z_j\|^2(q_n - 1) + \frac{E_{ij}^2 - \|z_i - Z_j\|^2(1 - q_n)}{2} \right) (z_j - \mu)(Z_j - \mu)^\top
\end{aligned}$$

CHAPTER 2. CMDS WITH PERTURBATION

If we denote

$$Y := \|z_i - Z_j\|E_{ij} - \|z_i - Z_j\|^2(q_n - 1) + \frac{E_{ij}^2 - \|z_i - Z_j\|^2(1 - q_n)}{2},$$

then we have

$$Y = \begin{cases} \|z_i - Z_j\|D_{ij} - \|z_i - Z_j\|^2(q_n - 1) + \frac{D_{ij}^2 - \|z_i - Z_j\|^2(1 - q_n)}{2}, & \text{with probability } 1 - q_n \\ -\|z_i - Z_j\|^2(q_n - 1) - \frac{\|z_i - Z_j\|^2(1 - q_n)}{2}, & \text{with probability } q_n \end{cases}$$

$$= \begin{cases} -\frac{q_n}{2}\|z_i - Z_j\|^2, & \text{with probability } 1 - q_n \\ -\frac{1 - q_n}{2}\|z_i - Z_j\|^2, & \text{with probability } q_n. \end{cases}$$

Hence, $\mathbb{E}[Y^2] = \frac{q_n(1 - q_n)}{4}\|z_i - Z_j\|^4$ and $\mathbb{E}[Y] = 0$.

Overall, we have the desired result. ■

Chapter 3

Manifold Denoising with Unsupervised Randomer Forest

3.1 Introduction

The accuracy, scalability, and applicability of many machine learning algorithms is currently impeded by the high-dimensional and large-scale nature of most modern data sets. In particular, the dimensionality, or number of features, of many data sets is often high, usually due to noise in the data – each data point is represented as a high-dimensional vector, but only a subset of them actually carries signals for subsequent inference. In other words, the data may live near some unknown low-dimensional manifold embedded in some high-dimensional space. To gain a thorough understanding of the data, it is therefore often necessary to reduce its dimensionality

CHAPTER 3. DENOISING WITH URERF

in a way that preserves its underlying structure. *Manifold learning* is a set of tools designed to recover the underlying latent low-dimensional manifold structures of high-dimensional data.

Existing manifold learning methods, however, face a number of challenges. Linear approaches such as principal component analysis (PCA) Pearson [1901], independent component analysis (ICA) Hyvärinen and Oja [2000], canonical correlation analysis (CCA) Hotelling [1936], multidimensional scaling (MDS) Cox and Cox [2000], CUR decompositions Mahoney and Drineas [2009], and Fisher’s linear discriminant analysis (LDA), have been widely applied and useful in many domains, but make fairly strong assumptions of about a linear structure underlying a data set. To mitigate these issues a number of methods that can be characterized as kernel PCA methods were devised Schölkopf et al. [1997], including Isomap Tenenbaum et al. [2000], Laplacian eigenmaps Belkin and Niyogi [2002], maximum variance unfolding Weinberger and Saul [2006]. These approaches are quite fragile to algorithm parameters, and typically require $\mathcal{O}(n^3)$ operations for n samples, which is prohibitively computationally expensive for many datasets. Methods based on exact nearest neighbors, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) Maaten and Hinton [2008], and Uniform Manifold Approximation and Projection (UMAP) McInnes and Healy [2018] also suffer computationally for large n . Approximate nearest neighbor approaches can mitigate some of these computational issues. For example, Fast Approximate Nearest-Neighbor Matching (FLANN) Muja and Lowe [2014] is a popular

CHAPTER 3. DENOISING WITH URERF

algorithm for nearest-neighbor detection in high-dimensional data sets. But FLANN, like all the above mentioned manifold learning algorithms, always operates on the observed dimensionality of the data. When the true manifold is low-dimensional, and the data are high-dimensional, the additional noise dimensions will be problematic for any of these algorithms.

We there propose an approach that we dub *Unsupervised Randomer Forest* (URerF). Unlike the previously described methods, URerF does not need to compute geodesic distances between pairs of points. Instead, URerF examines local structure by recursively clustering data in a sparse linear subspace of the original data, building on the recently proposed randomer forest algorithm for supervised learning Tomita et al. [2015]. This randomer forest approach allows URerF to separate meaningful structure in the data from the noise dimensions.

Another contribution of this manuscript is a novel method for evaluating manifold learning algorithms. Most existing manuscripts on the topic either embed the data into some low-dimensional space, such as 2D or 3D, and then merely visualize the results. This approach is obviously limited in a number of ways: (1) it is purely qualitative, (2) when the structure is higher dimensional it may be lost, and (3) it relies on an embedding, which introduces additional complications. Other manuscripts compare the results on some subsequent inference task, such as classification. Such an approach is only able to evaluate performance of the manifold learning algorithm composed with a particular subsequent inferential method, but not the manifold learning

algorithm itself. We therefore introduce Precision@K (or Recall@K) curves as quantitative metrics to evaluate manifold learning. The difference between our proposed metrics and standard metrics, is that we do not evaluate nearest neighbors with respect to the high-dimensional observed data, but rather the true low-dimensional latent representations. If a manifold learning does poorly on this metric, it has no hope to perform well on subsequent tasks. Indeed Precision@k provides a theoretical bound on subsequent classification accuracy Devroye et al. [1997].

3.2 Related Work

Of the many previously proposed manifold learning algorithms, we describe a few in detail that have risen to prominence and widespread use, which we will later compare to URerF.

One of the most widely used methods for nonlinear dimensionality reduction is still Isomap Tenenbaum et al. [2000], which constructs a low-dimensional embedding of input data by first finding geodesic distances between data points in a k-nearest neighbor graph, then applies classical multidimensional scaling to the matrix of graph distances. In the case of many noisy dimensions, however, the Isomap algorithm can fail to construct an accurate nearest-neighbor graph, and requires the storage of all point-to-point graph distances, which incurs both a large space and time complexity.

UMAP is a new algorithm for dimensionality reduction that efficiently reduces

CHAPTER 3. DENOISING WITH URERF

high-dimensional data to a low dimension using a fuzzy simplicial set representation of the input data points McInnes and Healy [2018]. Like other nearest-neighbor based algorithms, UMAP first constructs an undirected, weighted k-nearest neighbor graph from the input data, then embeds data points in a low-dimensional space using a force directed layout algorithm. The number of neighbors used to construct the graph in effect determines the local manifold structure that is to be preserved in the low-dimensional layout. In the force-directed layout approach, attractive forces between close vertices are iteratively balanced with repulsive forces between vertices that are far apart in the graph until convergence. UMAP builds upon the popular t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm, which attempts to preserve original interpoint distances in a much lower dimensional space. The Kullback-Leibler between the distribution of neighbor distances in the higher and lower dimensional spaces is used to determine the optimal mapping of points into the lower-dimensional space. t-SNE is primarily used for the visualization of high-dimensional data Maaten and Hinton [2008], and cannot be used with non-metric distances. The UMAP algorithm was shown to produce similar embeddings to t-SNE in two or three dimensions, but to scale better in terms of runtime across a wide range of embedding dimensions McInnes and Healy [2018].

Finally, most closely related to our method are existing unsupervised random forest methods, the most popular of which is included in Adele Cutler’s RandomForest R package Shi and Horvath [2006]. It proceeds by generating a synthetic copy of the

data by randomly permuting each feature independently of the others, and then attempts to classify the real versus the synthetic dataset. As will be seen below, this approach leads to missing surprisingly easy latent structures.

3.3 Unsupervised Randomer Forests

Our unsupervised random forest algorithm is based on the original Random Forest algorithm invented by Breiman Breiman [2001] with a few key distinctions. A random forest is an ensemble of decision trees where each tree is created from bootstrapped samples of the training dataset in order to incorporate randomness. That is, each tree is built from a random subset of training data. More formally, a random forest is as a classifier consisting of a collection of tree structured classifiers $\{h(\mathbf{x}, \theta_k)\}, k = 1, 2, \dots, \text{numtrees}$ where θ_k 's are parameters and each tree casts a unit vote for the most popular class at input \mathbf{x} .

Random Forests can be used for both supervised and unsupervised situations. In particular, for the unsupervised case, we can build a similarity matrix from the random forest. The similarity between two data points X_1 and X_2 is estimated as the fraction of the trees in which X_1 and X_2 appear in the same leaf node.

Our method additionally differs from the traditional, supervised case in the following ways. First, we describe three different splitting criteria to rank potential splits, inspired by techniques studied in clustering research. We also show empirically

CHAPTER 3. DENOISING WITH URERF

that a novel FastBIC split results in better precision and recall accuracy because we incorporate model selection at each node. In past literature, these splitting criteria have not been explored in conjunction with decision trees and random forests.

Second, we use the term *randomer* to label our technique, as our splitting methods are in addition based on random sparse linear combination of features to further improve randomness, as in Tomita et al.’s *Randomer Forests* algorithm Tomita et al. [2015]. Randomness is therefore incorporated into the forest generation process in candidate dimension generation stage as well as the traditionally randomness-inducing bagging stage.

Third, we correctly implement a previously proposed method for generating proximity matrices from random forests. Specifically, in the currently most widely used implementation of Random Forest Liaw and Wiener [2002], the aggregated normalized proximity matrices of N Random Forests with M trees each is not stochastically equivalent to the aggregated normalized proximity matrices of M Random Forests with N trees each. We fix this bug in our implementation.

3.4 Algorithm

3.4.1 Overall algorithm

Given an input data set $X = \{x_1, \dots, x_n\}$, T decision trees are built, each from a random sample of size $m < n$. In each tree, URerF recursively splits a parent node

CHAPTER 3. DENOISING WITH URERF

into its two child nodes. Each internal node bisects the data based on its value in a particular dimension or on its values in a linear combination of a small number of dimensions. The best dimension, or combination of dimensions, is selected to split on based on the score that results from the splitting criteria described in Section 3.4.2. The proximity matrix is then populated by computing the fraction of the trees in which every pair of elements reside in the same leaf node. Algorithm 1 describes the algorithm to build the forest and Algorithm 2 describes the tree building process. Algorithms 3 - 5 describe the splitting procedures. All algorithms are relegated to Section 3.4.4.

3.4.2 Splitting Criteria

We have implemented and compared several splitting criteria in our evaluation of the unsupervised random forest algorithm. Namely, we compare splitting by 2-means clustering, two-means clustering with the Bayesian Information Criterion (BIC) test, and a soft clustering as defined by the most likely Gaussian Mixture Model (GMM) of two Gaussians with the BIC test. The advantage of incorporating the BIC test into our splitting mechanism is its ability to select the split which results in two distinct clusters. The BIC test outputs a score which is a measure of how well the datapoints are explained by a Gaussian mixture model with two Gaussians.

3.4.2.1 Two-Means Splitting

We aim to find the split which minimizes the sum of the intra-cluster variance on the projected dimension. The elements are first sorted. Each element is scanned to be a potential cut point. The element to the left form one cluster with mean μ_1 and the elements to the right of the cutpoint form another cluster with mean μ_2 . We seek to find the cutpoint which minimizes the one-dimensional 2-means objective

$$\sum_{i=1}^{N_1} (X_i - \mu_1)^2 + \sum_{i=N_1}^{N_1+N_2} (X_i - \mu_2)^2.$$

The algorithmic details are explained in Algorithm 3.

3.4.2.2 Two-Means Splitting with FastBIC

The Bayesian Information Criterion (BIC) is used to select from among a finite collection of models. It is based on the log likelihood of the model given the points, with a regularization term penalizing complex models with many parameters. Concretely, the BIC score (y) can be defined as follows.

$$y = \ln(n)d - 2\ln(\widehat{L}) \tag{3.1}$$

Here \widehat{L} is the maximum log likelihood function of a particular model M , n is the sample size (number of data points) and d is the number of parameters estimated by

CHAPTER 3. DENOISING WITH URERF

the model. The maximum likelihood is defined as $\hat{L} = p(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the parameters that maximize the likelihood function, and x is the observed data, for the model M .

We rank potential splits on the BIC score obtained assuming a model with a mixture of two Gaussians; the sorted elements along a given dimension to the left of the cut point belong to one Gaussian and the elements to the right belong to another Gaussian. Note here, that this is slightly different from a Gaussian mixture model because here we assume that the point has been generated from each Gaussian k with a fixed (known) probability π_k which is the fraction of data points on either end of the split.

For example, if the elements in one dimension are $\{1, 3, 4, 6\}$, then the possible splits are $\{\{1\}, \{3, 4, 6\}\}$, $\{\{1, 3\}, \{4, 6\}\}$, and $\{\{1, 3, 4\}, \{6\}\}$. In the case $\{\{1, 3\}, \{4, 6\}\}$, we assume the model to comprise of two univariate Gaussians where $\{1, 3\}$ are sampled from the first and $\{4, 6\}$ are sampled from the second. $\pi_1 = \pi_2 = 0.5$ in this case, as an equal number of points are sampled from both Gaussians. We compute the estimates μ_1, μ_2 and σ_1^2, σ_2^2 for the means and variances of each of the two Gaussians in order to find the likelihood of the data under this split, and use it to compute the BIC score. The BIC score is computed for each possible split and we assign this dimension to the split which results in the lowest BIC score. We proceed in the same manner for all dimensions and choose to split using the dimension which results in the lowest overall BIC score. The model can be defined as follows. $P(x_1, \dots, x_n | \mu, \sigma, \pi)$

CHAPTER 3. DENOISING WITH URERF

is defined as follows, where $\mu = (\mu_1, \dots, \mu_K)$, $\sigma = (\sigma_1, \dots, \sigma_K)$, $\pi = (\pi_1, \dots, \pi_K)$:

$$P(x_{1:N}, z_{1:N,1:K} | \mu_{1:K}, \sigma, \pi) = \prod_{n=1}^N \prod_{k=1}^K \{\pi_k \mathcal{N}(x_n | \mu_k, \sigma_k^2)\}^{z_{n,k}} \quad (3.2)$$

Here N is the number of observations and K is the number of Gaussian clusters. In this case $K = 2$. We assume the points to the left of the split point form one cluster and the points to right form the second cluster. $z_i \in \{0, 1\}^K$ is the indicator vector for data point x_i : $z_{(i \in (0, s], k=0)} = 1$ and $z_{(i \in [s+1, N), k=1)} = 1$ and 0 otherwise. Now, letting the split point be x_s , $x_1 \dots x_s$ belong to cluster 1 and $x_{s+1} \dots x_N$ belong to cluster 2. The likelihood can be obtained by summing over the z s as follows.

$$P(x | \mu, \sigma, \pi) = \sum_{z_{1:N,1:K}} \prod_{n=1}^N \prod_{k=1}^K \{\pi_k \mathcal{N}(x_n | \mu_k, \sigma_k^2)\}^{z_{n,k}} \quad (3.3)$$

Equation (3.3) can be simplified by noting that $z_{(i \in (0, s], k=0)} = 1$ and $z_{(i \in [s+1, N), k=1)} = 1$ and 0 otherwise.

$$P(x | \mu, \sigma, \pi) = \prod_{n=1}^s \pi_1 \mathcal{N}(x_n | \mu_1, \sigma_1^2) \prod_{n=s+1}^N \pi_2 \mathcal{N}(x_n | \mu_2, \sigma_2^2) \quad (3.4)$$

The maximum log likelihood function

$\hat{L} = \log P(x_1, x_2, \dots, x_N | \hat{\mu}_{1:K}, \hat{\sigma}_{1:k}, \hat{\pi}_{1:k})$ is given by

$$\hat{L} = \sum_{n=1}^s [\log \pi_1 + \log \mathcal{N}(x_n | \hat{\mu}_1, \hat{\sigma}_1^2)] + \sum_{n=s+1}^N [\log \pi_2 + \log \mathcal{N}(x_n | \hat{\mu}_2, \hat{\sigma}_2^2)] \quad (3.5)$$

CHAPTER 3. DENOISING WITH URERF

Here, $\hat{\sigma}_k$ and $\hat{\mu}_k$ are the estimates of σ_k and μ_k which maximize the log likelihood function. Let $N_1 = s$ and $N_2 = N - (s + 1)$. Then, $\mu_k = \frac{1}{N_k} \sum_{n=1}^{N_k} x_n$, $\hat{\sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \|x_n - \hat{\mu}_k\|^2$, and $\hat{\pi}_k = \frac{N_k}{N}$. We further test for the single variance case ($\sigma_1 = \sigma_2$) and use the BIC formula to determine the best case. Substituting into equation 3.5 and simplifying, we get the following expression for the log likelihood.

$$\hat{L} = -\frac{N_1}{2} \log 2\hat{\pi}\hat{\sigma}_1^2 - \frac{N_2}{2} \log 2\hat{\pi}\hat{\sigma}_2^2, \quad (3.6)$$

where we have dropped a number of terms that are not functions of the parameters.

This FastBIC procedure is, to our knowledge, novel.

3.4.2.3 2-GMM Splitting with BIC

The GMM likelihood equations are the same as in the above case, except we now relax the binary constraints on the z_i . Specifically, each point is considered to belong to weighted sum of the two Gaussians as opposed to a single Gaussian. We do not know the values of z_i and we iteratively estimate them using the Expectation Maximization algorithm. Thus, whereas FastBIC is guaranteed to obtain the global maximum likelihood estimator, BIC is liable to find only a local maximum. The algorithm to estimate μ_k , π_k and z_k is provided in Algorithm 4.

3.4.3 Proximity Matrix Construction

The proximity matrix S for input data $D \in \mathbb{R}^{n \times d}$ is estimated using the unsupervised random forest by simply counting the fraction of times that a pair of points occurs in the same leaf node in the forest. Thus $S(i, j) = S_{ij} = \frac{L_{ij}}{T_{ij}}$, where $L(i, j)$ is the number of occurrences of points i and j in the same leaf node, and T_{ij} is the number of trees in which both point i and point j were included in the sample R that was used to build the tree.

3.4.4 Algorithms

Algorithm 1 Build a random forest using unlabeled data

```

1: procedure BUILDURF( $X, T, d, c$ )
2:   for  $i = 1, 2, \dots, T$  do
3:      $S \leftarrow$  random sample (with replacement) from  $X$  of size  $m$ 
4:      $t_i = \text{BuildTree}(S, d, c, 0)$ 
5:      $F \leftarrow F \cup \{t_i\}$ 
6:   return  $F$ 
7: end for
8: end procedure

```

3.5 Simulations on Synthetic Datasets

We compare and contrast the performance of URerF, Isomap and UMAP in this section through simulated data examples. Note that each of the three algorithms generates a distance matrix Δ as a by product, which naturally give rise to the

Algorithm 2 Build an unsupervised decision tree

```

1: procedure BUILDTREE( $Y, d, c, k, \text{depth}$ )
2:   if  $|Y| \leq c$  OR  $\text{depth} == d$  then
3:     return LeafNode( $Y, \text{depth}$ ) ▷ Create a leaf node
4:   else
5:      $\mathcal{C} \leftarrow$  random sample of size  $k$  from  $\{1, \dots, d\}$ 
6:      $\text{min\_dists} \leftarrow \infty$ 
7:     for  $i \in \mathcal{C}$  do
8:        $Y^{(i)} \leftarrow Y[:, i]$ 
9:        $(\text{midpt}, \text{sum\_sq\_dists}) = \mathbf{OneD}(Y^{(i)})$ 
10:      if  $(\text{sum\_sq\_dists} < \text{min\_dists})$  then
11:         $\text{best\_dim} = i$ 
12:         $\text{best\_split\_pt} = \text{midpt}$ 
13:      end if
14:    end for
15:     $Y_{\text{left}} = \{y \in Y | y(\text{best\_dim}) < \text{best\_split\_pt}\}$ 
16:     $Y_{\text{right}} = \{y \in Y | y(\text{best\_dim}) \geq \text{best\_split\_pt}\}$ 
17:     $\text{new\_Node} = \text{CreateInternalNode}(\text{best\_split\_pt}, \text{best\_dim}, \text{depth})$ 
18:     $\text{new\_Node.leftChild} = \text{BuildTree}(Y_{\text{left}}, d, c, k, \text{depth}+1)$ 
19:     $\text{new\_Node.rightChild} = \text{BuildTree}(Y_{\text{right}}, d, c, k, \text{depth}+1)$ 
20:    return new_Node
21:  end if
22: end procedure

```

Algorithm 3 Find the optimal split, in terms of the k -means objective, of one-dimensional data with $k=2$

```

1: procedure UNIVARIATETWOMEANS( $Z = \{z | z \in \mathcal{R}^1\}$ )
2:    $\mu_1 \leftarrow \min_{z \in Z} z$ 
3:    $\mathcal{C}_1 \leftarrow \{\mu_1\}$ 
4:    $\mathcal{C}_2 \leftarrow Z \setminus \mathcal{C}_1$ 
5:    $\mu_2 \leftarrow \frac{1}{|\mathcal{C}_2|} \sum_{z_i \in \mathcal{C}_2} z_i$  ▷ mean of  $\mathcal{C}_2$ 
6:    $\text{dists\_sq} \leftarrow \sum_{j=1}^2 \sum_{z_i \in \text{set}_j} (z_i - \mu_j)^2$ 
7:    $\text{min\_dist\_sq} \leftarrow \text{dists\_sq}$ 
8:   while  $\text{set}_2 \neq \emptyset$  do
9:      $z \leftarrow \min(\text{set}_2)$ 
10:     $\mathcal{C}_1 \leftarrow \mathcal{C}_1 \cup \{z\}$ 
11:     $\mathcal{C}_2 \leftarrow \mathcal{C}_2 \setminus \{z\}$ 
12:     $\mu_1 \leftarrow \frac{1}{|\mathcal{C}_1|} \sum_{z_i \in \mathcal{C}_1} z_i$  ▷ mean of  $\mathcal{C}_1$ 
13:     $\mu_2 \leftarrow \frac{1}{|\mathcal{C}_2|} \sum_{z_i \in \mathcal{C}_2} z_i$  ▷ mean of  $\mathcal{C}_2$ 
14:     $\text{dists\_sq} \leftarrow \sum_{j=1}^2 \sum_{z_i \in \mathcal{C}_j} (z_i - \mu_j)^2$ 
15:    if  $\text{dists\_sq} < \text{min\_dist\_sq}$  then
16:       $\text{min\_dists\_sq} \leftarrow \text{dists\_sq}$ 
17:       $\text{best\_midpt} \leftarrow (\max(\mathcal{C}_1) + \min(\mathcal{C}_2))/2$  ▷ Midpoint between  $\mathcal{C}_1$  and  $\mathcal{C}_2$ 
18:    end if
19:  end while
20:  return ( $\text{best\_midpt}, \text{min\_dist\_sq}$ )
21: end procedure

```

Algorithm 4 Find the optimal split, in terms of BIC score, of one-dimensional data with $k=2$ assumming the GMM Model

```

1: procedure GMM(Initialized estimates)
2:    $z_{n,k} = \frac{\mathcal{N}(x_n | \mu_k, \sigma^2) \pi_k}{\sum_{k'} \mathcal{N}(x_n | \mu_{k'}, \sigma^2) \pi_{k'}}$ 
3:    $\pi_k = \frac{1}{N} \sum_n z_{n,k}$ 
4:    $\mu_k = \frac{\sum_n z_{n,k}}{\sum_{n'} z_{n',k}}$ 
5:    $\sigma^2 = \frac{1}{N} \sum_n \sum_k z_{n,k} \|x_n - \mu_k\|^2$ 
6: end procedure

```

Algorithm 5 Find the optimal split, in terms of BIC score, of one-dimensional data with $k=2$

```

1: procedure ONED( $Z = \{z | z \in \mathcal{R}^1\}$ )
2:    $\mu_1 \leftarrow \min_{z \in Z} z$ 
3:    $\mathcal{C}_1 \leftarrow \{\mu_1\}$ 
4:    $\mathcal{C}_2 \leftarrow Z \setminus \mathcal{C}_1$ 
5:    $\mu_2 \leftarrow \frac{1}{|\mathcal{C}_2|} \sum_{z_i \in \mathcal{C}_2} z_i$  ▷ mean of  $\mathcal{C}_2$ 
6:    $\text{BIC\_curr} \leftarrow \sum_{j=1}^2 \sum_{z_i \in \text{set}_j} (z_i - \mu_j)^2$ 
7:    $\text{min\_BIC\_curr} \leftarrow \text{dists\_sq}$ 
8:   while  $\text{set}_2 \neq \emptyset$  do
9:      $z \leftarrow \min(\text{set}_2)$ 
10:     $\mathcal{C}_1 \leftarrow \mathcal{C}_1 \cup \{z\}$ 
11:     $\mathcal{C}_2 \leftarrow \mathcal{C}_2 \setminus \{z\}$ 
12:     $\mu_1 \leftarrow \frac{1}{|\mathcal{C}_1|} \sum_{z_i \in \mathcal{C}_1} z_i$  ▷ mean of  $\mathcal{C}_1$ 
13:     $\mu_2 \leftarrow \frac{1}{|\mathcal{C}_2|} \sum_{z_i \in \mathcal{C}_2} z_i$  ▷ mean of  $\mathcal{C}_2$ 
14:     $\sigma_1^2 \leftarrow \frac{1}{|\mathcal{C}_1|} \sum_{z_i \in \mathcal{C}_1} (z_i - \mu_1)^2$ 
15:     $\sigma_2^2 \leftarrow \frac{1}{|\mathcal{C}_2|} \sum_{z_i \in \mathcal{C}_2} (z_i - \mu_2)^2$ 
16:     $\sigma_{\text{comb}}^2 \leftarrow \frac{1}{|\mathcal{C}_1| + |\mathcal{C}_2|} \sum_{j=1}^2 \sum_{z_i \in \mathcal{C}_j} (z_i - \mu_j)^2$ 
17:     $\text{BIC\_diff\_var} \leftarrow -2(|\mathcal{C}_1| \log \frac{|\mathcal{C}_1|}{|\mathcal{C}_1| + |\mathcal{C}_2|} - \frac{|\mathcal{C}_1|}{2} \log 2\pi\sigma_1^2 - |\mathcal{C}_2| \log \frac{|\mathcal{C}_1|}{|\mathcal{C}_1| + |\mathcal{C}_2|} +$   

      $\frac{|\mathcal{C}_2|}{2} \log 2\pi\sigma_2^2) + \ln(3)(|\mathcal{C}_1| + |\mathcal{C}_2|)$ 
18:     $\text{BIC\_same\_var} \leftarrow -2(|\mathcal{C}_1| \log \frac{|\mathcal{C}_1|}{|\mathcal{C}_1| + |\mathcal{C}_2|} - \frac{|\mathcal{C}_1|}{2} \log 2\pi\sigma_{\text{comb}}^2 - |\mathcal{C}_2| \log \frac{|\mathcal{C}_1|}{|\mathcal{C}_1| + |\mathcal{C}_2|} +$   

      $\frac{|\mathcal{C}_2|}{2} \log 2\pi\sigma_{\text{comb}}^2) + \ln(2)(|\mathcal{C}_1| + |\mathcal{C}_2|)$ 
19:     $\text{BIC\_curr} \leftarrow \min(\text{BIC\_same\_var}, \text{BIC\_diff\_var})$ 
20:    if  $\text{BIC\_curr} < \text{min\_BIC}$  then
21:       $\text{min\_BIC} \leftarrow \text{BIC\_curr}$ 
22:       $\text{best\_midpt} \leftarrow (\max(\mathcal{C}_1) + \min(\mathcal{C}_2))/2$  ▷ Midpoint between  $\mathcal{C}_1$  and  $\mathcal{C}_2$ 
23:    end if
24:  end while
25:  return ( $\text{best\_midpt}, \text{min\_BIC}$ )
26: end procedure

```

CHAPTER 3. DENOISING WITH URERF

notion of nearest neighbors. These distance matrices Δ will be the main object we use to quantify the performance of a particular algorithm on a manifold dataset, using our modified “Precision@K” as evaluation criteria. Recall that Precision@K, Recall@K are defined in the following manner:

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|},$$

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}.$$

For a particular data point i in the dataset (corresponding to the i th row of distance matrix), letting the “relevant document” be the true K-nearest neighbor of i on the manifold, we can modify the above definitions to be:

$$(\text{Precision@K})_i = \frac{|\{\text{Knn of } i \text{ on the manifold}\} \cap \{\text{Knn of } i \text{ in the distance matrix } \Delta\}|}{K},$$

$$(\text{Recall@K})_i = \frac{|\{\text{Knn of } i \text{ on the manifold}\} \cap \{\text{Knn of } i \text{ in the distance matrix } \Delta\}|}{|\text{Knn of } i \text{ on the manifold}|},$$

Note that in this modified definition, $(\text{Precision@K})_i$ and $(\text{Recall@K})_i$ coincide with each other. Averaging over all rows in the distance matrices gives the Precision@K and Recall@K. Note that the underlying local manifold structure should ideally be preserved through higher precision scores and lower recall scores. One of the advantages of URerF is that under the presence of high dimensional noise, its performance does not degrade significantly when the dimension of noise increase. In the next

section, simulation on synthetic datasets are used to demonstrate this phenomenon.

3.5.1 Synthetic Datasets

Consider the following four datasets:

- **Linear:** wherein each point p is parameterized by $p = (4t, 6t, 9t)$, with $t \in (0, 1)$.
- **Helix:** wherein each point p is parameterized by $p = (2t \cos(2t), 2t \sin(2t), 2t)$, with $t \in (\frac{3}{2}\pi, \frac{9}{2}\pi)$.
- **Sphere:** wherein each point p is parameterized by $p = (r \cos(u) \sin(v), r \sin(u) \sin(v), r \cos(v))$, with $u \in (0, 2\pi)$, $v \in (0, \pi)$ and $r = 9$.
- **Gaussian Mixture:** wherein each point p is from a mixture of Gaussian distribution: $\sum_{i=1}^3 \pi_i \mathcal{N}(\mu_i, \Sigma_i)$, with $(\pi_1, \pi_2, \pi_3) = (0.3, 0.5, 0.2)$, $(\mu_1, \mu_2, \mu_3) = \left(\begin{bmatrix} -3 \\ -3 \\ -3 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} \right)$ and $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

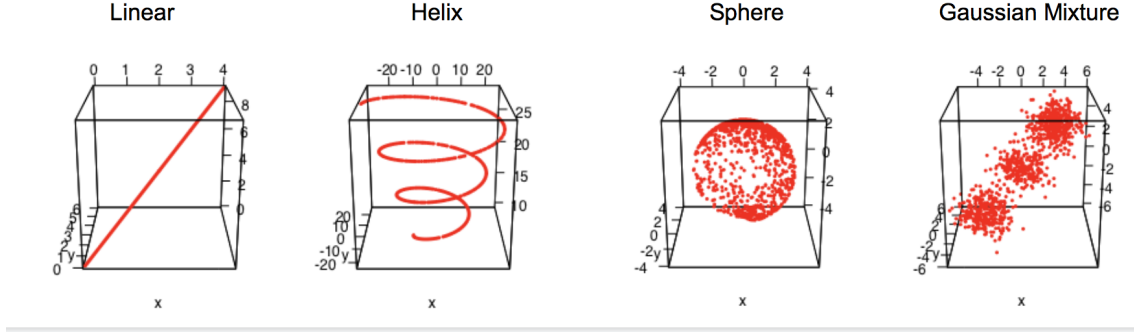


Figure 3.1: Synthetic Datasets

For any two data points p_i and p_j on the above manifolds with the given parameterization, geodesic distances between them $D_{geo}(p_i, p_j)$ can be easily calculate:

- **Linear:** $D_{geo}(p_i, p_j) = \sqrt{(4(t_i - t_j))^2 + (6(t_i - t_j))^2 + (9(t_i - t_j))^2}$
- **Helix:** $D_{geo}(p_i, p_j) = \int_{t_i}^{t_j} \sqrt{((2 \cos(2x) - 4p \sin(2x))^2 + (2 \sin(2x) + 4x \cos(2x))^2 + 4)} dx$
- **Sphere:** $D_{geo}(p_i, p_j) = r \arccos\left(\frac{p_i}{\|p_i\|} \cdot \frac{p_j}{\|p_j\|}\right)$
- **Gaussian Mixture:** We simply take the Euclidean distance between each pair of point as a surrogate for geodesic distance between them.

One thousands points are uniformly sampled from each of the above datasets (points from **Gaussian Mixture** dataset will be sampled according to the specified distribution). Of interests are the following four distance matrices calculated on these 1000 data points: Geodesic, URerF, Isomap and UMAP. Geodesic distance matrix will serve as the bench mark to which every other method is compared.

To see URerF is robust to high dimensional noise, Gaussian noises with varying

CHAPTER 3. DENOISING WITH URERF

dimensions d' are concatenated onto the synthetic datasets to form new datasets with noise for testing the performance of each algorithm. Specifically, for each data point $X_i \in \mathbb{R}^d$, generated noise $Y_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, cI) \in \mathbb{R}^{d'}$ are concatenated onto X_i , where c is some large constant and I is the $d' \times d'$ identity matrix, thus obtaining data points with noise: $Z_i = [X_i^\top | Y_i^\top]^\top \in \mathbb{R}^{d+d'}$. URerF, Isomap and UMAP distance matrices are then imputed, and compared with Geodesic distance matrix to obtain the Geodesic Precision@K. For $c = 70$ and $d = 2, 4, 6, 8, 10$, Figure 3.2 shows the performance of each algorithm.

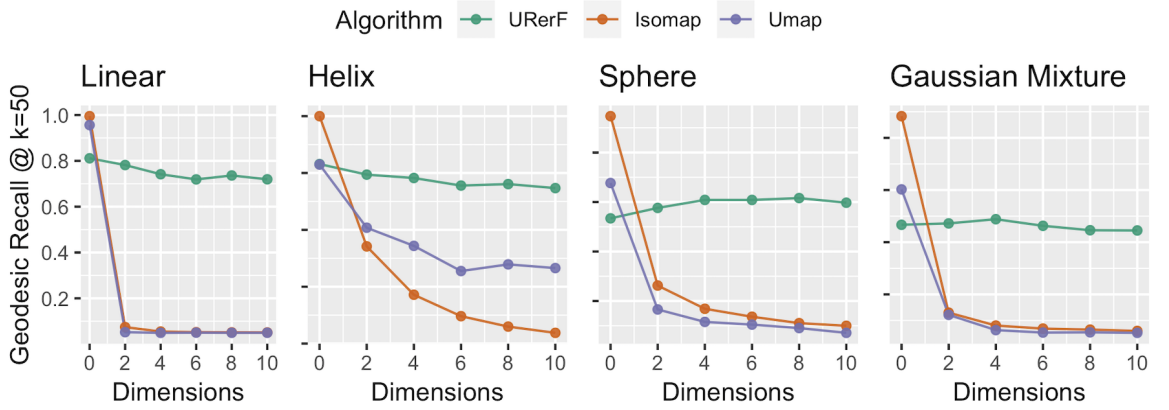


Figure 3.2: Geodesic Precision at K=50 with varying noise dimension d' .

Chapter 4

Seeded Graph Matching Via Joint Optimization of Fidelity and Commensurability

4.1 Introduction

Given m graphs, the graph matching problem (GMP) seeks to find a set of correspondences (i.e., “matchings”) between the vertex sets that best preserves similar substructures across the graphs Garey and Johnson [1979]. Because of its practical applicability, there exist numerous approximate graph matching algorithms in the literature; for an excellent survey of the existing literature, see Conte et al. [2004]. When matching across m graphs, often partial correspondences, or seedings, between

CHAPTER 4. SGM WITH JOFC

the vertices of some pairs of graphs are known. One cutting-edge algorithm for seeded graph matching, the Seeded Graph Matching (SGM) algorithm of Fishkind et al. [2012] and Lyzinski et al. [2013], leverages the information contained in seeded vertices to efficiently match graphs with thousands of vertices, achieving excellent performance with relatively few seeds.

However, as demonstrated in Fishkind et al. [2012], SGM achieves its optimal performance in the case of highly structured simple graphs on identical vertex sets. Although it can be modified to handle directed, weighted, and other non-simple graphs, in the presence of these generalizations the performance of the SGM algorithm deteriorates.

Moreover, the algorithm cannot currently handle cases where the graphs have different numbers of vertices or matchings across graphs that are not one-to-one. Often graphs arising from real data contain many of the aforementioned characteristics, and more robust procedures are needed to effectively match these graphs. Herein we present a new seeded graph matching algorithm derived from the Joint Optimization of Fidelity and Commensurability (JOFC) algorithm of Priebe et al. [2013]. Our algorithm is flexible enough to handle many of the difficulties inherent to real data, while simultaneously not sacrificing too much performance (compared to SGM) when matching across simple graphs. The paper is laid out as follows: In Section 4.2, we define the classical GMP and present the details of the SGM algorithm. In Sections 4.3.2 – 4.3.4, we present our JOFC seeded graph matching problem in detail. In Sec-

tion 4.4, we match multiple time series graphs generated from a zebrafish brain, again demonstrating the validity of our JOFC algorithm (see Figure 4.2). Throughout this chapter, we will define $P(n)$ to be the set of $n \times n$ permutation matrices, and $D(n)$ to be the set of $n \times n$ doubly stochastic matrices.

4.2 The classical graph matching problem and the SGM algorithm

In its classical form, the two-graph matching problem is as follows: Given two graphs G_1 and G_2 on the same vertex set V (with $|V| = n$), we seek to find a bijection $\phi : V \mapsto V$ that minimizes the number of edge disagreements induced by ϕ ; specifically, we seek a bijection $\phi : V \mapsto V$ that minimizes

$$d(\phi) = |\{(i, j) \in V \times V : [i \sim_{G_1} j, \phi(i) \not\sim_{G_2} \phi(j)] \text{ or } [i \not\sim_{G_1} j, \phi(i) \sim_{G_2} \phi(j)]\}|. \quad (4.1)$$

Equivalently stated, if the adjacency matrices for G_1 and G_2 are respectively A and B , the problem seeks a permutation matrix $P \in P(n)$ minimizing $\|A - PBP^T\|_F$, where $\|\cdot\|_F$ is the usual matrix Froebenius norm. If we allow G_1 and G_2 to be directed, loopy and weighted, then the classical graph matching problem is equivalent to the quadratic assignment problem, and therefore is known to be NP-hard. Hence no

CHAPTER 4. SGM WITH JOFC

efficient exact graph matching algorithm is known.

We can generalize the above classical graph matching problem between two graphs to matchings among m graphs: Given m graphs G_1, G_2, \dots, G_m on the same vertex set V (with $|V| = n$), we seek to find a set of bijections $\Phi = \{\phi^{p,q}\}_{1 \leq p < q \leq m}$, where $\phi^{p,q} : V \mapsto V$ is a mapping between graph G_p and G_q , such that the set Φ minimizes

$$d(\Phi) = \sum_{\phi^{p,q} \in \Phi} |\{(i, j) \in V \times V : [i \sim_{G_p} j, \phi^{p,q}(i) \sim_{G_q} \phi^{p,q}(j)] \text{ or } [i \sim_{G_p} j, \phi^{p,q}(i) \not\sim_{G_q} \phi^{p,q}(j)]\}|. \quad (4.2)$$

Equivalently, if the adjacency matrix for graph G_k is A_k , the problem seeks a set of permutation matrices $\{P_{p,q}\} \subseteq P(n)$ minimizing $\sum_{1 \leq p < q \leq m} \|A_p - P_{p,q} A_q P_{p,q}^T\|_F$, where $\|\cdot\|_F$ is the usual matrix Froebenius norm.

Often when matching across graphs, we have access to a partial matching of the vertices in the form of seedings. If we are given, for each graph G_k a subset $S_k \subset V$ of size s_k of the vertices called *seeds* and *seeding* functions $\sigma^{p,q}$ which maps a subset of S_p to a subset of S_q , the classical seeded graph matching problem (SGMP) then seeks to minimize (4.1) over bijections $\phi^{p,q} : V \mapsto V$ satisfying $\phi^{S_p, S_q} = \sigma^{p,q}$ where ϕ^{S_p, S_q} is the restriction of $\phi^{p,q}$ to an appropriate subset of S_p .

For the two-graph matching case, the state-of-the-art approximate seeded graph matching algorithm, the SGM algorithm of Fishkind et al. [2012] and Lyzinski et al. [2013], begins by relaxing the SGMP to minimize $\|A - PBP^T\|_F$ over doubly stochas-

CHAPTER 4. SGM WITH JOFC

tic matrices P of the form $P = I_s \oplus P'$ with $P' \in D(n - s)$. The algorithm then utilizes Frank-Wolfe methodology to efficiently solve the relaxed problem and finally projects this relaxed solution onto $P(n)$. The deterioration of the SGM algorithm's performance on non-simple graphs motivates the need for more robust seeded graph matching procedures, such as the JOFC algorithm presented herein.

It bears noting that there are a multitude of variations on the classical problem in the literature, where different graph attributes give rise to objectives other than minimizing (4.1). For an excellent survey of the existing literature, including many of the current variations on the classical problem, see Conte et al. [2004]. We choose here to focus on the classical problem (4.1), as it is closely related to the SGM seeded graph matching algorithm.

4.3 Seeded graph matching via JOFC

We presently approach the seeded graph matching problem via a modification of the Joint Optimization of Fidelity and Commensurability (JOFC) algorithm of Priebe et al. [2013], which was originally designed for manifold matching. Briefly, our algorithm embeds a list of m graphs into a common Euclidean space where our matching inference task can be performed. The embedding seeks to maximize the information contained both within the connectivity structure of each graph and the across graph relationship provided by the seeding, i.e. we seek to maximize the f_i -

CHAPTER 4. SGM WITH JOFC

delity and *commensurability* of the embedding. Once embedded, finding the optimal matching between the vertices then amounts to solving a generalized tensor assignment problem.

We will present our algorithm in its most general form and will note when certain assumptions on our graphs necessarily lead to simplifications.

4.3.1 Setup

Let $\{G_k\}_{k=1}^m$ be graphs on respective vertex sets $\{V_k := V(G_k)\}_{k=1}^m$ (with each $|V_k| = n_k$). Without loss of generality, we assume the vertices are labeled $V_k = \{1, 2, \dots, n_k\}$. As we no longer assume all graphs have the same set of vertices, the graph matching problem as stated in (4.1) is not necessarily well-posed. Rather than reformulating the classical GMP in terms general enough to handle all of the difficulties inherent to real data problems, we choose instead to reformulate our approach to graph matching. We begin with a set of assumed true matching $\mathcal{M} = \{M^{p,q}\}_{1 \leq p < q \leq m}$ between every pair of vertex sets V_p and V_q , though in the present general setting a matching is simply a subset $M^{p,q} \subset V_p \times V_q$. If $(u, v) \in M^{p,q}$, then u and v are “matched” vertices in graph G_p and G_q , though the precise definition of “matched” here is context specific. In one setting, u and v could be the same actor in two different communication graphs. In another setting, u and v could represent the same neuron in two different neuro-connectome graphs. In each of our real and simulated data examples, the true matching is explicit from the context of the problem. Note

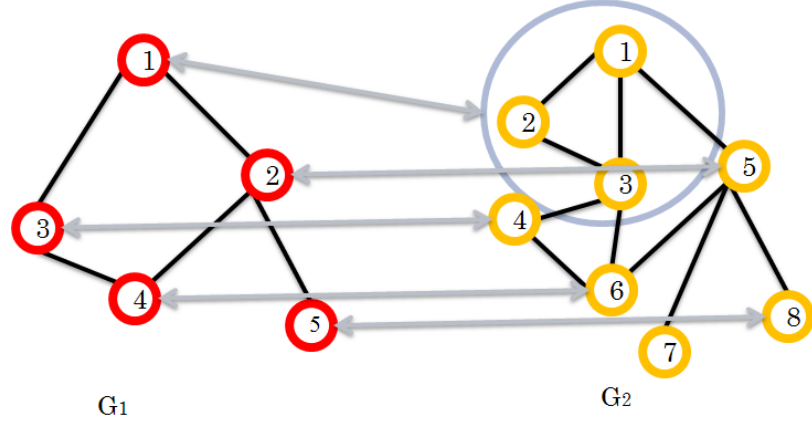


Figure 4.1: For graphs G_1 and G_2 , the matching $M_{1,2} \subset V_1 \times V_2$ is given via the gray arrows and is formally defined as $M^{1,2} = \{(1, 1), (1, 2), (1, 3), (2, 5), (3, 4), (4, 6), (5, 8)\}$. Note that vertex 1 in G_1 is matched to three vertices in G_2 , namely vertices 1, 2, and 3. Also vertex 7 in G_2 is not matched to any vertex in G_1 .

that although it is often the case that $n_p = n_q$ and the matching $M^{p,q}$ is a bijection between the vertex sets, our more general definition allows for multiple vertices in G_p to be matched to a single vertex or no vertex at all in G_q , and vice-versa. See Figure 4.1 for an illustrative example of a matching in this general context. Our inference task is then to leverage the information contained in seeded vertices to estimate the true underlying matching \mathcal{M} .

In this newly reformulated graph matching problem, a *seeding* refers to a collection $\mathcal{S} = \{S^{p,q}\}_{1 \leq p < q \leq m}$, where given any two graphs G_p and G_q , $S^{p,q} \subset M^{p,q} \subset V_p \times V_q$ with the following property: if $(i, j) \in S^{p,q}$, $(l, h) \in S^{p,q}$ and $(i, h) \notin S^{p,q}$ then $(i, h) \notin M^{p,q}$; i.e. seeded vertices can only be matched to other seeded vertices if the matching is explicitly given by the seeding. This is an intuitive assumption; indeed it is often natural to assume the full matchedness amongst the seeded vertices is known.

CHAPTER 4. SGM WITH JOFC

There are certainly applications in which this is not true, and our algorithm can easily be modified to incorporate an incomplete seeding as well. The vertices in the ordered pairs of $S^{p,q}$ are referred to as *seeds*, and we define $S_{p,q}$ as the set of all seeds in V_p that matches to seeds in V_q . Furthermore, we define all seeded elements in V_p as S_p , i.e., $S_p := \bigcup_{q:q \neq p} S_{p,q} = \bigcup_{q:q \neq p} \{i : (i,j) \in S^{p,q}\}$. We will label all the unseeded vertices of G_p and G_q via $U_{p,q}$ and $U_{q,p}$ respectively, and define $U_p = V_p \setminus S_p$, in addition, we will assume also that $(i,j) \notin M^{p,q}$ for $i \in S_{p,q}$, $j \in U_{q,p}$ or $i \in U_{p,q}$, $j \in S_{q,p}$.

Note: To simplify later notation, for any subset $W \subset V_p \times V_q$, we write

$$W(i) := \begin{cases} \{j \in V_q : (i,j) \in W\} & \text{if } i \in V_p \\ \{j \in V_p : (j,i) \in W\} & \text{if } i \in V_q. \end{cases}$$

In the sequel, we shall also write $|S_p| = s_p$.

4.3.2 Embedding the seeded graph data

Due to the pathological nature of the graphs we aim to match, performing the matching directly on the graph data proved difficult. To circumvent this, our algorithm uses multidimensional scaling (MDS) to embed the m graphs into a common Euclidean space where our matching task can more readily occur.

Our embedding begins with m dissimilarity representations $\{\Delta_k \in \mathbb{R}^{n_k \times n_k} : k = 1, \dots, m\}$ of $\{G_k\}_{k=1}^m$. We will assume a priori that the dissimilarities have been nor-

CHAPTER 4. SGM WITH JOFC

malized to be on the same scale. Ideally, we choose the dissimilarity dependent on the nature of the data, as different dissimilarities will emphasize different aspects of the underlying graph topology. Although we do not theoretically address the issue of optimally choosing the dissimilarity in the present paper, empirical results have shown that correctly choosing the dissimilarities is essential to the performance of our downstream matching task. Indeed, in one application to matching neural connectomes of the *C. elegans* worm, we achieve excellent performance using the weighted DICE dissimilarity of Angelelli et al. [2008], a local neighborhood based measure suitable for the sparse structure of the worm brain graphs. However, in the simulated Erdős-Rényi (ER) graph examples, the DICE dissimilarity is not appropriate due to the highly structured nature of the neighborhoods in ER graphs. We empirically demonstrate a marked performance increase by utilizing a more global dissimilarity, namely the shortest path distance. Alternately, we could have used diffusion distance, expected commute times, etc. See Yen et al., Bunke and Riesen [2008], Pekalska and Duin [2005] for a wealth of possible dissimilarity representations.

In order for the matching inference task to successfully occur in the embedded space, the embedding must preserve the information contained both within the connectivity structure of each graph and the between graph relationship given by the matching \mathcal{M} . In essence, the goal of the embedding is simple: If $(i, j) \in M^{p,q}$, then $i \in G_p$ and $j \in G_q$ are “matched” vertices and should be embedded close to each other in \mathbb{R}^d . Also, if $u, v \in V_i$ are such that $\Delta_i(u, v)$ is small, then u and v are similar

CHAPTER 4. SGM WITH JOFC

vertices in the underlying graph and should also be embedded close to each other in \mathbb{R}^d .

Preserving the matching \mathcal{M} in the MDS embedding (or preserving any available across graph relationship) requires us to impute an across graph dissimilarity $\delta^{p,q} : V_p \times V_q \mapsto \mathbb{R}$. For matched vertices $(i, j) \in M^{p,q}$, it is reasonable to impute $\delta^{p,q}(i, j) = 0$, though for $(i, j) \notin M$, the imputation is less obvious. Here we treat these dissimilarities as missing data in the subsequent MDS procedure.

We do not have access to the full matching \mathcal{M} , but the seeding \mathcal{S} provides sufficient information for calculating the imputed δ amongst the seeded vertices. For $(i, j) \in S^{p,q}$, we (as before) impute $\delta^{p,q}(i, j) = 0$. For $(i, j) \in N^{p,q} := \{(i, j) \in S_p \times S_q : (i, j) \notin S^{p,q}\}$, we take $\delta^{p,q}(i, j)$ as missing data. Furthermore, rather than incur additional estimation error by imputing the unknown δ across the unseeded vertices, we also treat these as missing data in our MDS procedure.

We proceed then as follows. We first embed the seeded vertices and then out-of-sample embed the unseeded vertices using the methodology of Tang et al. [2013]. With a possible relabeling of the vertices, let the seeded vertices for each graph G_p be denoted as $S_p = \{1, 2, \dots, s_p\}$, so that

$$\Delta_p = \begin{matrix} & \begin{matrix} S_p & U_p \end{matrix} \\ \begin{matrix} S_p \\ U_p \end{matrix} & \begin{pmatrix} \Delta_{1,1}^{(p)} & \Delta_{1,2}^{(p)} \\ \Delta_{2,1}^{(p)} & \Delta_{2,2}^{(p)} \end{pmatrix} \end{matrix}.$$

CHAPTER 4. SGM WITH JOFC

Labeling the embedded vertices of S_p via $\{X_1^{(p)}, X_2^{(p)}, \dots, X_{s_p}^{(p)}\}$, we define the across-graph squared *commensurability error* of the embedding between graph G_p and G_q via

$$\varepsilon_{C, G_p, G_q}^2 := \sum_{1 \leq i \leq s_p} \sum_{j \in S^{p,q}(i)} \left(d(X_i^{(p)}, X_j^{(q)}) - \delta^{p,q}(i, j) \right)^2, \quad (4.3)$$

and the total squared across-graphs *commensurability error* is given by:

$$\varepsilon_C^2 := \sum_{1 \leq p < q \leq m} \varepsilon_{C, G_p, G_q}^2 \quad (4.4)$$

where $d(\cdot, \cdot)$ is the Euclidean distance between points in \mathbb{R}^d . For $(i, j) \in S^{p,q}$, we impute $\delta^{p,q}(i, j) = 0$ and commensurability error between two graphs reduces simply to the squared Euclidean distance between embedded matched vertices. The commensurability error captures how well the embedding preserves the partial graph matching provided by the seeding.

Note that even if the commensurability of the embedding is small, the embedded points may poorly preserve the original within-graph dissimilarities, which is captured by the fidelity of our embedding. The within-graph squared *fidelity error* of the embedding of $\Delta_{1,1}^{(p)}$ is given by

$$\varepsilon_{F_p}^2 := \sum_{1 \leq i < j \leq s_p} \left(d(X_i^{(p)}, X_j^{(p)}) - \Delta_{1,1}^{(p)}(i, j) \right)^2, \quad (4.5)$$

CHAPTER 4. SGM WITH JOFC

and the total squared *fidelity error* is

$$\varepsilon_F^2 := \sum_{k=1}^m \varepsilon_{F_k}^2. \quad (4.6)$$

Closely connected to the fidelity error is the across-graph squared *separability error* between two graphs defined via

$$\varepsilon_{S, G_p, G_q}^2 := \sum_{(i,j) \in (S_p \times S_q \setminus S^{p,q})} \left(d(X_i^{(p)}, X_j^{(q)}) - \delta^{p,q}(i, j) \right)^2. \quad (4.7)$$

and the total across-graphs *separability error* is

$$\varepsilon_S^2 := \sum_{1 \leq p < q \leq m} \varepsilon_{S, G_p, G_q}^2 \quad (4.8)$$

However, since we took $\delta^{p,q}(i, j)$ as missing data in this scenario, we can ignore the separability error for later inference.

If the errors ε_F^2 , and ε_C^2 are all small and we have embedded the graphs into an appropriate dimension, then we can successfully perform our matching inference task in the target embedding space. Assuming at present that we know a suitable embedding dimension d , we simultaneously control the above errors by jointly embedding

CHAPTER 4. SGM WITH JOFC

the $\sum_{k=1}^m s_k$ seeded vertices of the m graphs via the *omnibus dissimilarity matrix*

$$D := \begin{matrix} & S_1 & S_2 & \dots & S_m \\ \begin{matrix} S_1 \\ S_2 \\ \vdots \\ S_m \end{matrix} & \begin{pmatrix} \Delta_{1,1}^{(1)} & \delta^{1,2} & \dots & \delta^{1,m} \\ \delta^{2,1} & \Delta_{1,1}^{(2)} & \dots & \delta^{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ \delta^{m,1} & \delta^{m,2} & \dots & \Delta_{1,1}^{(m)} \end{pmatrix} \end{matrix} \quad (4.9)$$

We embed D using the JOFC algorithm of de Leeuw [1977] for weighted raw stress MDS, where the associated weight matrix is given by

$$W := \begin{bmatrix} J_{s_1} - I_{s_1} & w \cdot O^{1,2} & \dots & w \cdot O^{1,m} \\ w \cdot O^{2,1} & J_{s_2} - I_{s_2} & \dots & w \cdot O^{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ w \cdot O^{m,1} & w \cdot O^{m,2} & \dots & J_{s_m} - I_{s_m} \end{bmatrix} \in \mathbb{R}^{(\sum_{k=1}^m s_k) \times (\sum_{k=1}^m s_k)},$$

where J_n is an $n \times n$ matrix of all entries being 1, w is a fixed real number between 0 and 1 representing the weight we choose, and $O^{p,q}$ is a matrix with the same dimensionality as that of $\delta^{p,q}$, whose entries take value 1 whenever the corresponding entries in $\delta^{p,q}$ is 0, and take value 0 whenever the corresponding value is missing in $\delta^{p,q}$ Lyzinski et al. [2016].

Suppose \mathbf{X} is some configurations of the $\sum_{k=1}^m s_k$ points in \mathbb{R}^d , the JOFC algorithm

CHAPTER 4. SGM WITH JOFC

is an iterative procedure minimizing the cost function

$$\sigma(\mathbf{X}) = \sum_{i < j} W_{i,j} (D(i, j) - d_{i,j}(\mathbf{X})) \quad (4.10)$$

over all possible configurations of $\sum_{k=1}^m s_k$ points \mathbf{X} in \mathbb{R}^d . The $W_{i,j}$ are the weights representing our confidence in the dissimilarity $D(\cdot, \cdot)$ between pairs of vertices. In our applications, W is designed so that (4.10) simplifies to

$$\sigma(\mathbf{X}) = w\varepsilon_F^2 + (1 - w)\varepsilon_C^2, \quad (4.11)$$

a mixture of the fidelity/separability errors (which capture how well our embedding preserves the original within-graph dissimilarities) and the commensurability error (which captures how well the embedding preserves the partial matching given by the seeds). This ability to weight the dissimilarities is an essential feature of the JOFC algorithm and is one of the main reasons we have chosen it over more classical multidimensional scaling procedures. In all of our applications, we have chosen $w = 0.8$, and have left the optimal choice of w for future work.

4.3.3 Embedding the unseeded vertices

We next use the procedures outlined in Tang et al. [2013] to out-of-sample embed all the unseeded vertices U_k $k = 1, \dots, m$, obtaining the configuration $\mathbf{Y} \subset \mathbb{R}^d$ of

CHAPTER 4. SGM WITH JOFC

the $u_k = |U_k|$ unseeded vertices of each graph G_k , labeled $\{Y_1^{(k)}, \dots, Y_{u_k}^{(k)}\}$. For the out-of-sample embedding, we treat the unknown across-graph dissimilarities involving unseeded vertices as missing data.

The goal of our out-of-sample procedure is simply to preserve the within graph dissimilarities $\Delta_{1,2}^{(k)}$'s. Indeed, suppose that $(i, j) \in U_p \times U_q$ is such that $(i, j) \in M^{p,q}$. Ideally, the seeding $S^{p,q}$ will be such that there exists $(u, v) \in S^{p,q}$ such that $\Delta_p(i, u)$ and $\Delta_q(j, v)$ will both be small. If our two step embedding procedure preserves the seeding $S^{p,q}$ and $\Delta_{1,1}^{(p)}$, $\Delta_{1,1}^{(q)}$, $\Delta_{1,2}^{(p)}$, and $\Delta_{1,2}^{(q)}$ then $d(Y_i^{(p)}, Y_j^{(q)})$ will be small from a simple triangle inequality argument. If $(i, j) \in U_p \times U_q$ is such that $(i, j) \notin M^{p,q}$, then the seeding $S^{p,q}$ ideally has the property that there exists $(u, v) \in S^{p,q}$ such that one of $\Delta_p(i, u)$ and $\Delta_q(j, v)$ is small and the other one is large. If our two step embedding procedure preserves the seeding $S^{p,q}$ and $\delta^{p,q}$, $\Delta_{1,1}^{(p)}$, $\Delta_{1,1}^{(q)}$, $\Delta_{1,2}^{(p)}$, and $\Delta_{1,2}^{(q)}$ then $d(Y_i^{(p)}, Y_j^{(q)})$ will be large from another simple triangle inequality argument. Assuming the above, the matching M amongst unseeded vertices will then be preserved under the embedding without the need to impute the unknown δ across unseeded vertices.

Following Tang et al. [2013], our embedding procedure then seeks to minimize

CHAPTER 4. SGM WITH JOFC

the stress function:

$$\sigma(\mathbf{Y}) = \sum_{p=1}^m \sum_{i=1}^{s_p} \sum_{j=1}^{u_p} \tilde{w}^{p,q}(X_i^{(p)}, Y_j^{(p)}) \left(d(X_i^{(p)}, Y_j^{(p)}) - \Delta_{1,2}^{(p)}(i, j) \right)^2 \quad (4.12)$$

$$+ \sum_{p=1}^m \sum_{i=1}^{u_p} \sum_{j=1}^{u_p} \tilde{w}^{p,q}(Y_i^{(p)}, Y_j^{(p)}) \left(d(Y_i^{(p)}, Y_j^{(p)}) - \Delta_{2,2}^{(p)}(i, j) \right)^2 \quad (4.13)$$

over configurations \mathbf{Y} . Here $\tilde{w}^{p,q}(\cdot, \cdot) : V_p \times V_q \mapsto \mathbb{R}$ is a weighting function representing our confidence in the computed dissimilarity between pairs of vertices. In our applications, we have chosen to zero out the weighting function $\tilde{w}^{p,q}$ between unseeded vertices within each graph, i.e. we have zeroed out the sums in (4.13) from $\sigma(\mathbf{Y})$. We set the remaining \tilde{w} 's to be 1. This is an artifact of our implementation of the out-of-sample embedding procedure, and is not a requirement of our algorithm. However, in applications where only the 1-neighborhoods of the seeded vertices are known, this would be a naturally enforced constraint.

4.3.4 Matching the unseeded vertices

If the Euclidean distances amongst the unseeded vertices well preserves the unknown matching M (i.e. if $(i, j) \in U_{p,q} \times U_{q,p}$ is in $M^{p,q}$, then $d(Y_i^{(p)}, Y_j^{(q)})$ is small and if $(i, j) \notin M^{p,q}$, then $d(Y_i^{(p)}, Y_j^{(q)})$ is large), then we approximate the unknown matching M between unseeded vertices as follows:

CHAPTER 4. SGM WITH JOFC

1. Match graph $H = G_1$ and G_2 by solving the generalized assignment problem:

$$\min_{\widehat{M^{1,2}} \subset U^{1,2}} \sum_{(i,j) \in \widehat{M^{1,2}}} d(Y_i^{(1)}, Y_j^{(2)}) \quad (4.14)$$

To avoid trivial solutions, we impose the further restriction that $|\widehat{M^{1,2}}(j)| > 0$ for all $j \in U_1$. We do allow vertices in U_2 to be unmatched to any vertices in U_1

2. Averaging the matched graphs we get from G_1 and G_2 to get a new graph H , that is, if k_1 vertices of graph G_1 is matched with k_2 vertices of graph G_2 , we construct the corresponding vertex in H as the average of these $k_1 + k_2$ vertices.
3. Match graph G_3 to this new graph H while enforcing the consistency. That is, if vertex i in G_1 matches to both vertex j in G_2 and vertex l of G_3 in the seeding, then j and l should also match.
4. Get the new H by taking the average of H and G_3 after matching, as in ii.
5. Repeat this process for every graph G_k , $k = 4, \dots, m$.

We can thus match every G_k to the rest of the graphs while keep the seeding consistency. The generalized assignment problem is known to be NP-hard, see Catrysse and Van Wassenhove [1992] for background. However, there are many good polynomial-time approximation algorithms in the literature, see for example Shmoys and Tardos [1993], which we use in our examples.

CHAPTER 4. SGM WITH JOFC

In the case of matching only two graphs, letting \widehat{M} be our algorithm's approximation of the true matching M , we measure our performance via the *matched ratio* of the remaining unseeded vertices,

$$\delta^{(s_1+s_2)} = \frac{\left| \{i \in U_1 : \widehat{M}(i) \neq M(i)\} \right| + \left| \{j \in U_2 : \widehat{M}(j) \neq M(j)\} \right|}{u_1 + u_2}. \quad (4.15)$$

We measure the performance of graph matching algorithms by calculating the fraction, R_m , of the unseeded vertices correctly matched across the graphs. In the case where $s_1 = s_2 = m$ and $n_1 = n_2 = n$, we calculate

$$R_m = \frac{|\{i \in U_1 : \widehat{M}(i) \neq M(i)\}|}{n - m}. \quad (4.16)$$

When $s_1 \neq s_2$ and $n_1 \neq n_2$, R_{s_1, s_2} is calculated via

$$R_{s_1, s_2} = \frac{|\{i \in U_1 : \widehat{M}(i) \neq M(i)\}| + |\{i \in U_2 : \widehat{M}(i) \neq M(i)\}|}{u_1 + u_2}.$$

Note that the number of unseeded vertices to match decreases as the number of seeded vertices increases. In all examples, we show how increasing the number of seeded vertices from 0 to some substantive fraction of the total number of vertices significantly increases our relative performance in correctly matching the unseeded vertices.

4.4 Application: Matching zebrafish brains

We have a time series dataset collected by Prevedel et al. on zebrafish brain using Light-Field Deconvolution Microscopy and pan-neuronal expression of GCaMP, which is a fluorescent calcium indicator that serves as a proxy for neuronal activity Lyzinski et al. [2016], Prevedel et al. [2014]. We generate a time series of 20 graphs $\{G^{(t)}\}$ indexed by time $t = 1, 2, \dots, 20$ from this dataset in which each graph $G^{(t)}$ is generated on the same fixed 469 neurons (vertices) and the vertex i and j are adjacent if and only if the neuronal activities between neuron i and j cross some threshold at time t . That is, for each graph $G^{(t)}$, there is an associated adjacency matrix $A^{(t)}$, such that $A_{i,j}^{(t)} = 1$ if there is an edge between vertex i and j . We have, therefore, a complete true matching between vertices of all 20 graphs. Furthermore, the i th row of the adjacency matrix $A_{i,j}^{(t)}$ completely characterized the status of i th neuron in relation to other neurons at time t in the brain. We will demonstrate our JOFC scheme by comparing the matchings of the unseeded vertices with this known "ground truth". Initial change point detection analysis reveals that there was an anomaly occurring at $t = 13$. For our purpose, we select graphs $G^{(8)}, G^{(9)}, G^{(10)}, G^{(11)}, G^{(12)}, G^{(13)}$ and choose for each graph, the (same) first 300 vertices as seeds and out-of-sample the rest 169 vertices for matching. Since the anomaly occurs at $t = 13$, the first 5 graphs should match almost perfectly with each other while $G^{(13)}$ will not be correctly matched to any one of the first 5 graphs. As mentioned before, the choice of distance measure plays an crucial role in the graph matching problem, here we pick $1 - jaccard$

as our notion of distance Adamic and Adar [2003].

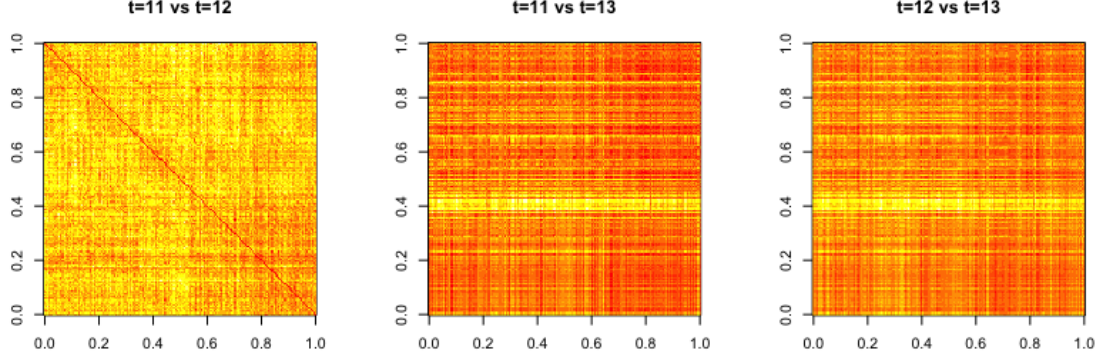


Figure 4.2: The pairwise distance heat map of the unseeded vertices from 3 graphs, at time $t=11$, $t=12$, and $t=13$. The darker colors corresponding to smaller distances and the X axis represents the unseeded vertices 301–469 in one graph and Y axis represents the unseeded vertices 469–301 in another graph. Since we know the true matchings between all vertices, we should ideally see a dark line from upper left to lower right between $t=11$ and $t=12$, which can be readily seen. However, since there is an anomaly at $t = 13$, the unseeded vertices at $t=13$ will not be matched correctly, which also can be seen.

To match the unseeded vertices, we calculate the pairwise distances between all pairs of unseeded vertices from different graphs and rank them in increasing order after the embedding. For example, between graphs $G^{(16)}$ and $G^{(17)}$, the k th unseeded vertex in $G^{(16)}$ will be more likely to be matched with the unseeded vertex in graph $G^{(17)}$ that is the closest to it. In this manner, not only we have a matching, but we also get a list of likely matchings between unseeded vertices.

Chapter 5

Discussion

5.1 Discussion on CLT for CMDS

In Athreya et al. [2016] and Levin et al. [2017], the authors prove that adjacency spectral embedding of the random dot product graph gives rise to a central limit theorem for the estimated latent positions. In this work we extend these results to the previously unexplored area of perturbation analysis for CMDS, addressing a gap in the literature as acknowledged in Fan et al. [2018] and Peterfreund and Gavish [2018]. Notably, the three noise models we proposed in Section 2.2.1 each give rise to a central limit theorem; that is, for Euclidean distance matrix, the rows of the configuration matrix given by CMDS under noise will center around the corresponding rows of the true configuration matrix. Furthermore, our simulations on the synthetic data together with the shape clustering data all demonstrated the validity of our

CHAPTER 5. DISCUSSION

results. We have avoided any discussion of the model selection problem of choosing a suitable embedding dimension \hat{d} . Instead, we assume d is known – except in Section 4.2. There are many methods for choosing (spectral) embedding dimensions, see Chatterjee [2015], Jackson [1991], Zhu and Ghodsi [2006].

One natural question can be raised is how to estimate the σ in the noise model of interests. However, we would like to point out that for our embedding method and associated theoretical results, consistent estimation of σ is not important. Indeed, the classical multidimensional scaling algorithm does not require estimating σ , but rather the dimension d of the original data points (see the description of classical multidimensional scaling in section 2.1). Under all of our noise model, $\|\mathbf{E}\| \leq \sigma\sqrt{n}$ and provided that we choose d such that $\lambda_d > n^{1/2+\epsilon}$ for any $\epsilon > 0$, then our theoretical limit results apply. For concreteness, we can choose $\epsilon = 1/3$ and thus as long as we choose the embedding dimension \hat{d} satisfying $\lambda_{\hat{d}}(B) \geq n^{2/3}$, then $\hat{d} \rightarrow d$ almost surely and our central limit theorem applies.

Throught this paper, we assume that d is fixed as $n \rightarrow \infty$. Therefore, given a central limit theorem for the embedding into d dimension, one can derive a central limit theorem for the embedding into $d' < d$ dimension in a straightforward manner. More specifically, given a dissimilarity matrix $\hat{\Delta}^{(2)}$ and positive integers $d' \leq d$, the classical multidimensional scaling of $\hat{D}^{(2)}$ into $\mathbb{R}^{d'}$ is equivalent to the classical multidimensional scaling of $\hat{\Delta}^{(2)}$ into \mathbb{R}^d and keeping the first $d' < d$ columns (see the description of classical multidimensional scaling in Section 1.1). Thus, our limit re-

CHAPTER 5. DISCUSSION

sults can be rephrased to say that, letting $\widehat{X}_n^{(d')}$ denote the classical multidimensional scaling of $\widehat{D}^{(2)}$ into $\mathbb{R}^{d'}$ for $d' < d$, that there exists a sequence of $d' \times d'$ orthogonal matrix $W_n^{(d')}$ and a sequence of $d \times d'$ matrices with orthonormal columns T_n such that

$$\sqrt{n} \left((\widehat{X}_n^{(d')} W_n^{(d')})_i - T_n (Z_n - \bar{Z}_n)_i \right)$$

converges to a mixture of multivariate normal. For a given n , T_n is a matrix corresponding the principal component projection of Z_n into $\mathbb{R}^{d'}$. We emphasize that T_n is not necessarily unique (indeed, the eigenvalues of the covariance matrix for Z_n are not necessarily distinct).

We further note that the dependency on d in our limit results is implicit in the covariance matrices. Naively speaking, we can say that the estimation accuracy is inversely proportional to d . This is most visible in the statement of Equation (1) (which is also a corollary of our results), since as d increases r also increases, note that $r \leq d + 2$. A more precise description is that the accuracy of our limit results depends on the covariance matrix Σ , which is a $d \times d$ matrix. Since the squared norm of a mean 0 multivariate Gaussian is the trace of its covariance matrix, we see that as d increases, the trace of Σ does not have to increase with d . Indeed, the trace of Σ depends purely on the distribution F of the underlying data points; in the case where the data points are sampled from a multivariate normal with mean 0 and identity matrix in \mathbb{R}^d , then as d increases, the trace of Σ also increases linearly.

Our presentation emphasizes the central limit theorem mainly because it is a suc-

CHAPTER 5. DISCUSSION

cinct limit results. Nevertheless, the uniform or global error bounds can be established in a similar manner. More specifically, the central limit theorem for a fixed index i is a consequence of applying the Lindeberg-Feller central limit theorem to Eq.(5) (which is a sum of independent mean 0 random variables). If, instead of the Lindeberg-Feller central limit theorem, we apply a concentration inequality a la Hoeffding/Bernstein, then we can show that for any index i , $\|(\hat{X}_n W_n)_i - (Z_i - \bar{Z})\| \leq Cn^{-1/2}$ with high probability. A union bound over the n rows of X_n then implies

$$\sup_{i \in [n]} \|(\hat{X}_n W_n)_i - (Z_i - \bar{Z})\| \leq C\sqrt{\frac{\log n}{n}}; \quad n^{-1} \sum_i \|(\hat{X}_n W_n)_i - (Z_i - \bar{Z})\| \leq C\sqrt{\frac{\log n}{n}}.$$

A practically relevant and conceptually illustrative example comes from relaxing the assumption of common variance for the entries of the noise matrix E in Section 2.2.2: the consistency result from Theorem 2.4.3 no longer holds. To illustrate this point, we return to our three-point-mass simulation presented in Section 2.5.1 and modify our noise model as follows: Let $\tilde{E}_{ij} \stackrel{i.i.d.}{\sim} \text{Uniform}(-D_{ij}, +D_{ij})$ for $i < j$ and $\tilde{E}_{ij} = \tilde{E}_{ji}$. (The noise now depends on the entries of D , and $\Delta = D + \tilde{E}$ no longer has negative entries.) The embedding of Δ into two dimensions gives class-conditional Gaussians; however, we have introduced bias into the embedding configuration. Figure 5.1 shows, for one realization, the embedding result. Note that the empirical mean and the theoretical positions do not coincide in simulation with large n , and theoretically even in the limit.

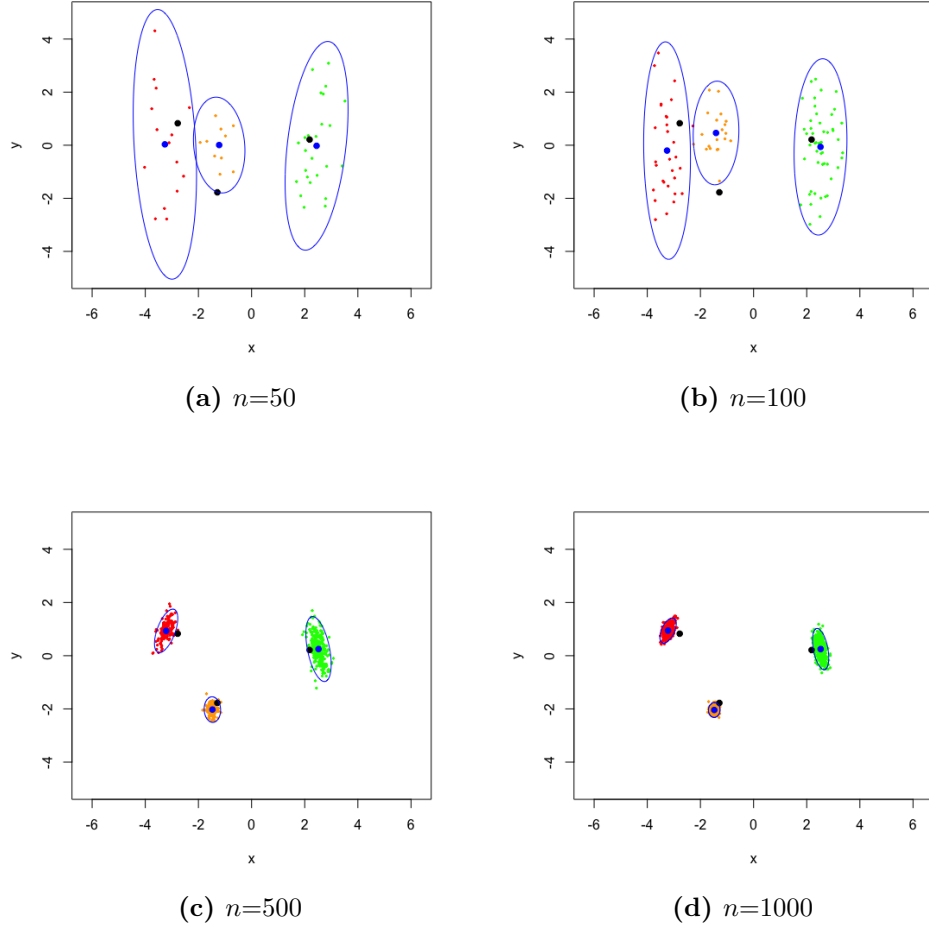


Figure 5.1: Simulation of CMDs with heteroscedastic noise \tilde{E} . The black dots are the true positions for the three points. The blue dots are the empirical means and the blue ellipses are the 95% level curve of the empirical covariance matrix. Note that \tilde{E} used in this simulation is of the same order for the off-diagonal blocks as that used in Figure 2.1. NB: there is asymptotic bias.

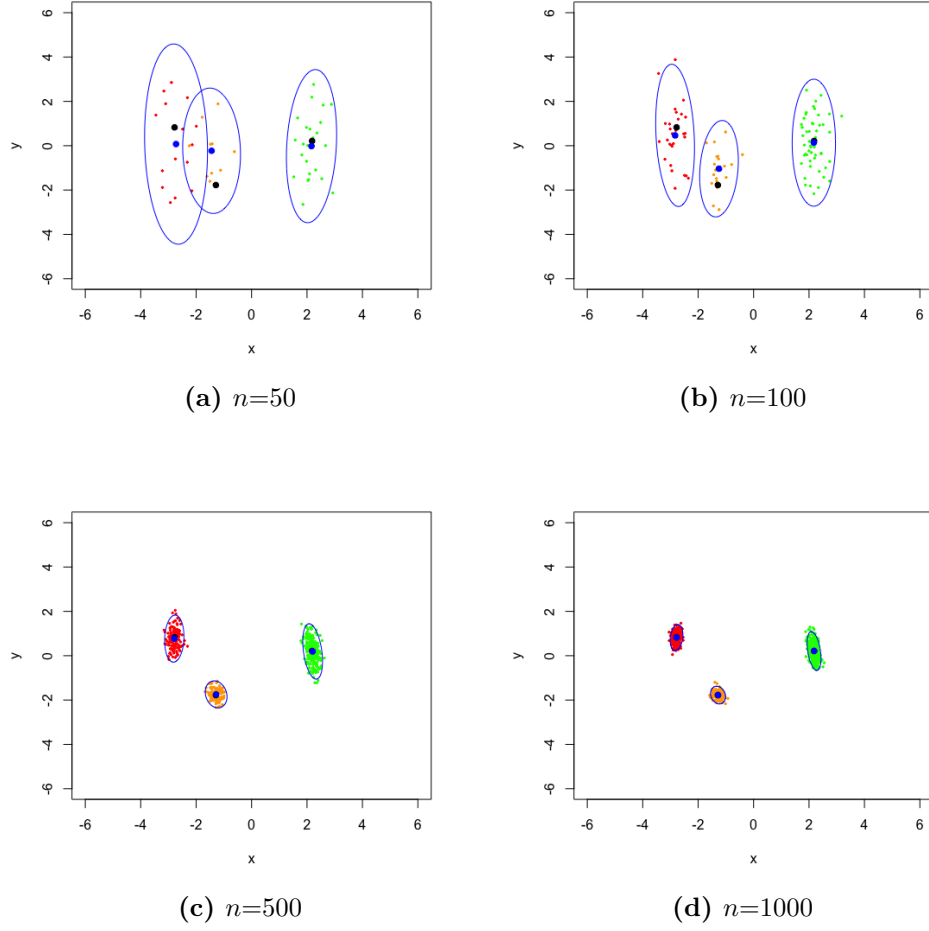


Figure 5.2: Simulation of MDS using raw stress criterion for $n=50$, 100, 500 and 1000 points. The black dots are the true positions of x_1 , x_2 and x_3 , the blue dots are the empirical mean of the simulation and the blue ellipses are the 95% level curve of the empirical covariance matrix.

CHAPTER 5. DISCUSSION

CMDS is just one of a wide variety of multidimensional scaling techniques. Minimizing the raw stress criterion is another commonly used MDS technique [de Leeuw and Heiser, 1982], i.e., given a $n \times n$ observed dissimilarity matrix Δ and an embedding dimension d , one seeks to minimize the objective function

$$\sigma_r = \sigma_r(X) = \sum_{(i,j)} (\delta_{ij} - \|X_i - X_j\|)^2.$$

The minimization of $\sigma_r(X)$ is with respect to all configurations $X \in \mathbb{R}^{n \times d}$ and usually proceeds via an iterative algorithm which updates the configuration matrix X until a stopping criterion is met. Keeping the simulation settings as in Section 2.5.1, the resulting configuration is shown in Figure 5.2. This suggests that the CLT may hold for raw stress just as well as for CMDS. However, this claim is at best a conjecture at present as perturbation analysis of stress minimization algorithms is significantly more involved.

5.2 Conjecture: CMDS on Omni Embedding of graphs and Hypothesis Testing

5.2.1 Problem Formulation

Given a collection of Random Dot Product Graphs: $G^{(1)}, G^{(2)}, \dots, G^{(m)}$ and their corresponding adjacency matrices $A^{(1)}, A^{(2)}, \dots, A^{(m)}$, where the G_i 's are vertex-matched graphs on the same vertex set V , with $|V| = n$, Levin et al. [2017] seeks to jointly embed all m graphs into a common (prespecified) d -dimensional Euclidean space by embedding the $nm \times nm$ OMNI Matrix

$$\mathcal{M} := \begin{bmatrix} A^{(1)} & \frac{A^{(1)}+A^{(2)}}{2} & \frac{A^{(1)}+A^{(3)}}{2} & \dots & \frac{A^{(1)}+A^{(m)}}{2} \\ \frac{A^{(2)}+A^{(1)}}{2} & A^{(2)} & \frac{A^{(2)}+A^{(3)}}{2} & \dots & \frac{A^{(2)}+A^{(m)}}{2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \frac{A^{(m)}+A^{(1)}}{2} & \frac{A^{(m)}+A^{(2)}}{2} & \frac{A^{(m)}+A^{(3)}}{2} & \dots & A^{(m)} \end{bmatrix}.$$

Using the notation in Levin et al. [2017], let $S_{\mathcal{M}}$ represent the $d \times d$ matrix of top d eigenvalues of \mathcal{M} , ordered by magnitude, and let $U_{\mathcal{M}}$ be the $nm \times d$ -dimensional matrix of associated eigenvectors. Define the omnibus embedding, denoted $\text{OMNI}(\mathcal{M})$, by $U_{\mathcal{M}}S_{\mathcal{M}}^{1/2}$, note that $\text{OMNI}(\mathcal{M})$ produces m separate points in Euclidean space for

CHAPTER 5. DISCUSSION

each graph vertex, effectively one such point for each copy of the multiple graphs in our sample. Furthermore, denote

$$U_{\mathcal{M}} S_{\mathcal{M}}^{1/2} = \begin{bmatrix} \widehat{X}^{(1)} \\ \widehat{X}^{(2)} \\ \vdots \\ \widehat{X}^{(m)} \end{bmatrix} \quad (5.1)$$

where each $\widehat{X}^{(i)}$ is an $n \times d$ matrix representing the embedding of the i th random graph in d dimension.

As pointed out in Levin et al. [2017], the omnibus embedding introduces alignment between graphs by placing an average on the off-diagonal blocks of \mathcal{M} , merely considering a Frobenius norm difference between blocks of the omnibus embedding

$$\|\widehat{X}^{(i)} - \widehat{X}^{(j)}\|_F \quad (5.2)$$

without any Procrustes alignments provides meaningful discrimination between graphs with different latent positions. Specifically, we can construct an $m \times m$ distance matrix Δ given by $\Delta_{ij} := \|\widehat{X}^{(i)} - \widehat{X}^{(j)}\|_F$ as an estimation for the Frobenius distance matrix D between true latent positions $D_{ij} := \|X^{(i)} - X^{(j)}\|_F$, where $X^{(i)}$ is the true latent position for the i th graph. That is, $\Delta = D + E$ for some $m \times m$ noise matrix E . Note that the entries of E are correlated. Furthermore, $\mathbb{E}[E] = 0$ only when all the

CHAPTER 5. DISCUSSION

m graphs shares the same underlying true latent position. Those two observations make the proof of the following conjectures quite a bit more challenging.

A natural thing to do now is to perform classical multidimensional scaling (CMDS) on Δ and D (we now call this procedure **CMDS** \circ **OMNI**(\mathcal{M}, d, k) and compare the resulting configuration matrix $\widehat{\mathcal{G}} = \text{CMDS}(\Delta) \in \mathbb{R}^{m \times k}$ and $\mathcal{G} = \text{CMDS}(D) \in \mathbb{R}^{m \times k}$.

We propose the following conjectures:

Conjecture 5.2.1 (*Central Limit Theorem for the Rows of CMDS \circ OMNI Under the Null*)

Using the above notations, we propose that if all graphs are generated from the same RDPG probability matrix \mathbf{P} , then there exist a sequence of $m \times m$ orthogonal matrix W_n , such that

$$\lim_{n \rightarrow \infty} \mathbb{P}\{\sqrt{n}[(\widehat{\mathcal{G}}W_n)_i - (\mathcal{G})_i] \leq \alpha\} = \Phi(\alpha, \Sigma(z)) \quad (5.3)$$

where z 's depends on the true underlying distribution of the latent positions X of the graphs. Here $(A)_i$ represent the i th row of matrix A viewed as a column vector and Φ is the cumulative distribution function of a multivariate Gaussian.

Note that we can replace $(\mathcal{G})_i$ with 0 in this case.

Conjecture 5.2.2 (*Central Limit Theorem for the rows of CMDS \circ OMNI Under the Alternative*)

Using the above notations, we propose that if graphs are generated from L distinct

CHAPTER 5. DISCUSSION

RDPG probability matrices $\mathbf{P}_1, \dots, \mathbf{P}_L$, then there exist a sequence of $m \times m$ orthogonal matrix W_n , such that

$$\lim_{n \rightarrow \infty} \mathbb{P}\{\sqrt{n}[(\widehat{\mathcal{G}}W_n)_i - (\mathcal{G})_i - b_i] \leq \alpha\} = \int_{\text{supp} F} \Phi(\alpha, \Sigma(z)) dF(z) \quad (5.4)$$

$$= \sum_{j=1}^L \pi_j \Phi(\alpha, \Sigma(z)) \quad (5.5)$$

where z and π_i s depend on the true underlying distribution of the latent positions X of the graphs and b represents bias incurred during the OMNI step. Here $(A)_i$ represent the i th row of matrix A viewed as a column vector and Φ is the cumulative distribution function of a multivariate Gaussian.

5.2.2 Simulation for the Choice of Scaling Factor

The proposed conjectures can be a primary tool for testing if a given set of graphs are from the same latent models (given by the same \mathbf{P} matrix, see Levin et al. [2017]). In this section, we will elucidate the choice of the scaling factor (\sqrt{n}) in our Conjectures. Throughout this section, $A^{(1)}, A^{(2)}, \dots, A^{(m)}$ are the $n \times n$ binary adjacency matrices of vertex-aligned graphs on the same vertex set.

5.2.2.1 Under the Alternative

We demonstrate that our conjectured Central Limit Theorem (CLT) holds only when $n \rightarrow \infty$ instead of when $m \rightarrow \infty$.

CHAPTER 5. DISCUSSION

Consider a collection of graphs given by two different 2-block SBMs, where

$$\mathbf{B}_1 = \left[\begin{array}{c|c} 0.3 & 0.1 \\ \hline 0.1 & 0.7 \end{array} \right],$$

$$\mathbf{B}_2 = \left[\begin{array}{c|c} 0.6 & 0.2 \\ \hline 0.2 & 0.4 \end{array} \right],$$

and $\boldsymbol{\pi} = (0.8, 0.2)^\top$. Let

$$A^{(1)}, \dots, A^{(m_1)} \stackrel{i.i.d.}{\sim} \text{SBM}(\mathbf{B}_1, \boldsymbol{\pi}),$$

$$A^{(m_1+1)}, \dots, A^{(m)} \stackrel{i.i.d.}{\sim} \text{SBM}(\mathbf{B}_2, \boldsymbol{\pi}).$$

In this particular simulation, let $\frac{m_1}{m} = 0.6$. We perform the procedure

$$\text{CMDS} \circ \text{OMNI}(\mathcal{M}, d = 2, k = 2)$$

with parameters $(n, m) = (10, 500), (10, 1000), (100, 100)$. The following plots are the CMDS results of the Δ matrix defined in Section 5.2.1, in which each point represents a single graph. In addition, the red points represent graphs generated from \mathbf{B}_1 and the black points are those from \mathbf{B}_2 . Note that from the plot, for a fixed n , when $m \rightarrow \infty$, the central limit theorem does not kick in, while when n increases, the plots exhibits roughly a mixture of two Gaussians.

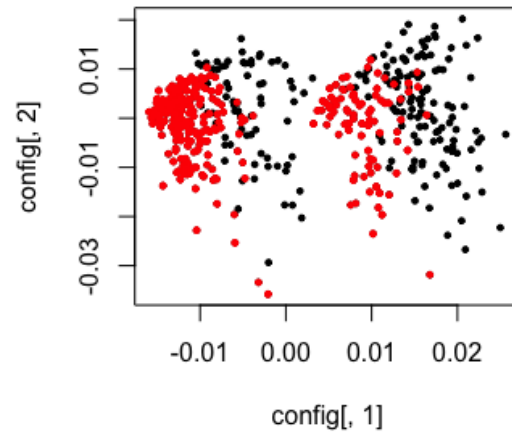


Figure 5.3: $n=10$, $m=500$

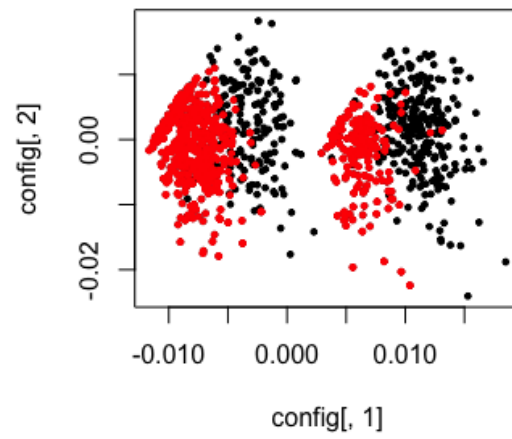


Figure 5.4: $n=10$, $m=1000$

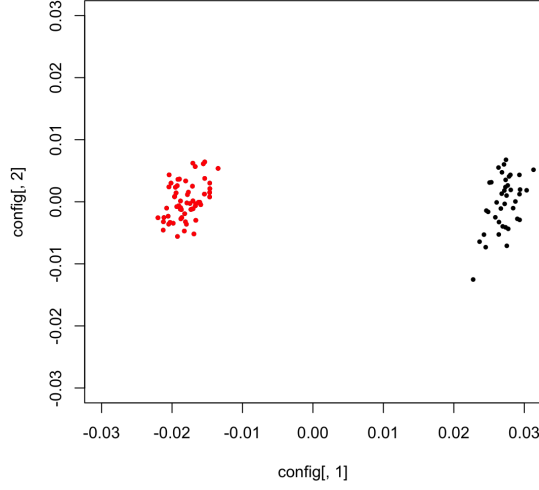


Figure 5.5: $n=100, m=100$

5.2.2.2 Under the Null

Now we demonstrate that the choice of scaling factor, \sqrt{n} in Conjecture 5.2.1 is reasonable. We generate graphs $A^{(1)}, \dots, A^{(m)} \stackrel{i.i.d.}{\sim} \mathbf{SBM}(\mathbf{B}_1, \boldsymbol{\pi})$, and let $g_1, \dots, g_m \in \mathbb{R}^2$ to be the cmds embedding results of the $\mathbf{CMDS} \circ \mathbf{OMNI}(\mathcal{M}, d=2, k=2)$. Let Q be the sample co-variance of g_i 's, we check to see if $\sqrt{n}Q$ remains reasonably unchanged when n is large. The parameters used in this simulation is $(n, m) = (500, 20), (1000, 20), (2000, 20)$. The following plots are the embedding of g_i 's, in which each point represents a graph.

CHAPTER 5. DISCUSSION

	$n=500$	$n=1000$	$n=2000$
Q	$\begin{bmatrix} 1.73 \cdot 10^{-5} & -2.96 \cdot 10^{-21} \\ -2.96 \cdot 10^{-21} & 1.69 \cdot 10^{-5} \end{bmatrix}$	$\begin{bmatrix} 8.37 \cdot 10^{-6} & -4.83 \cdot 10^{-22} \\ -4.83 \cdot 10^{-22} & 8.16 \cdot 10^{-6} \end{bmatrix}$	$\begin{bmatrix} 3.92 \cdot 10^{-6} & 1.02 \cdot 10^{-21} \\ 1.02 \cdot 10^{-21} & 3.89 \cdot 10^{-6} \end{bmatrix}$
$\sqrt{n}Q$	$\begin{bmatrix} 3.86 \cdot 10^{-4} & -6.61 \cdot 10^{-20} \\ -6.61 \cdot 10^{-20} & 3.78 \cdot 10^{-4} \end{bmatrix}$	$\begin{bmatrix} 2.65 \cdot 10^{-4} & -1.53 \cdot 10^{-20} \\ -1.53 \cdot 10^{-20} & 2.58 \cdot 10^{-4} \end{bmatrix}$	$\begin{bmatrix} 1.75 \cdot 10^{-4} & 4.57 \cdot 10^{-20} \\ 4.57 \cdot 10^{-20} & 1.74 \cdot 10^{-4} \end{bmatrix}$

Table 5.1: Sample Covariance and Scaled Sample Covariance Matrix

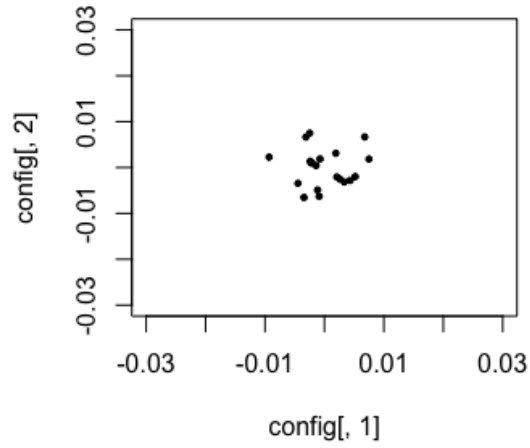


Figure 5.6: $n=500$, $m=20$

CHAPTER 5. DISCUSSION

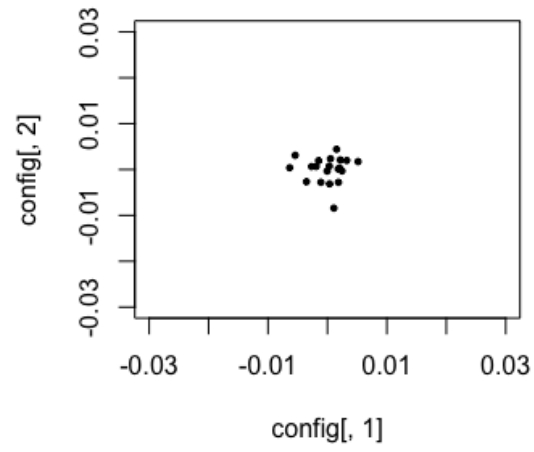


Figure 5.7: $n=1000$, $m=20$

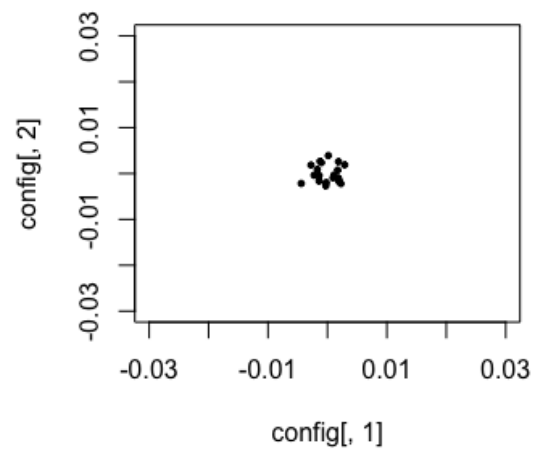


Figure 5.8: $n=2000$, $m=20$

5.2.3 Mouse Connectome

We have in our possession a collection of mouse connectome graphs collected on four different phenotypes, namely BTBR, C57, CAST and DB2. For each phenotype, we collect one graphs from each of the eight mice on the same 332 neurons (thus gives 32 vertex-matched mice connectome graphs). In this example, BTBR is discarded and only the rest three phenotypes are used for the purpose of demonstration of our methodology, that is, as our raw input data, we have 24 vertex-matched graphs $A^{(1)}, \dots, A^{(24)}$ with each graph having 332 vertex.

The raw graph data is pre-processed in the following manner:

1. **Centering** : $A^{(1)}, \dots, A^{(24)}$ are transformed to be centered around $\mathbf{0}$, so that the transformed matrices are $\tilde{A}^{(i)} = A^{(i)} - \bar{A}$, where $\bar{A} = \frac{1}{24} \sum_{i=1}^{24} A^{(i)}$.
2. **Pass to the Rank**: For each $\tilde{A}^{(i)}$, the entries in each row are ranked in decreasing order, and then each entry is replaced with its rank in that row (so that largest entry in each row will be replaced by m , etc.) to get a new matrix $\hat{A}^{(i)}$.
3. **Diagonal Augmentation**: Each diagonal entry of $\hat{A}^{(i)}$ is then replaced by the sum of the row in which that diagonal entry is located.

After preprocesing, the OMNI step is performed on the $\hat{A}^{(i)}$ matrices to get \mathcal{M} . Furthermore, an examination of the scree plot of the OMNI matrix \mathcal{M} in Figure 5.9 gives the graph embedding dimension $d = 3$.

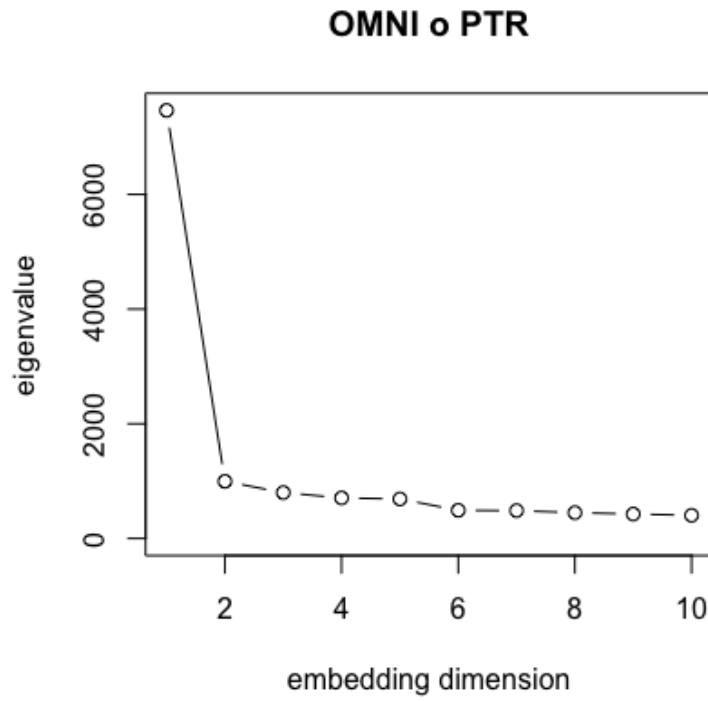


Figure 5.9: Embedding Dimension Selection of OMNI Matrix.

Dissimilarity matrix is then constructed from the OMNI embedding as described in equation 5.2, which is shown in Figure 5.10.

CHAPTER 5. DISCUSSION

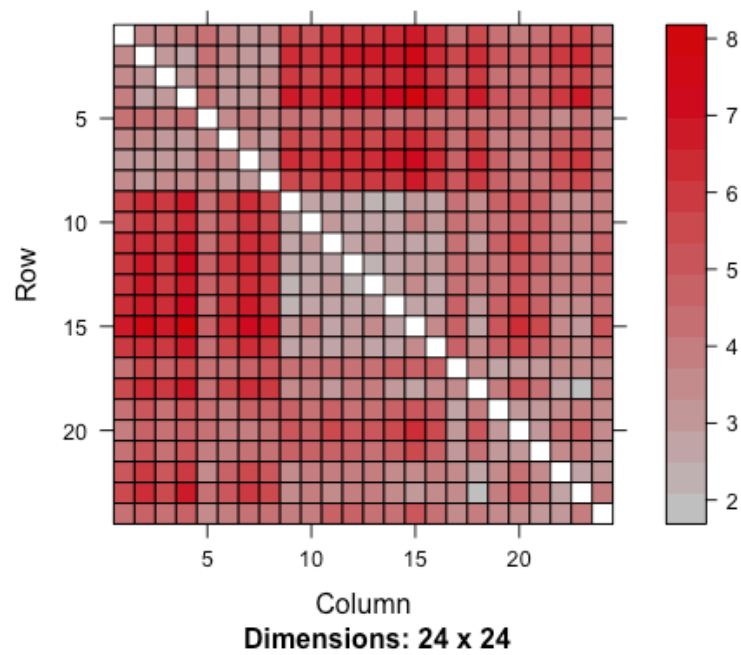


Figure 5.10: Dissimilarity Matrix Between Twenty-four Graphs.

In the final step, CMDS into embedding dimension $k = 2$ is performed on the 24×24 dissimilarity matrix, resulting the following embedding plot. Note that different phenotypes are well-separated in this embedding plot.

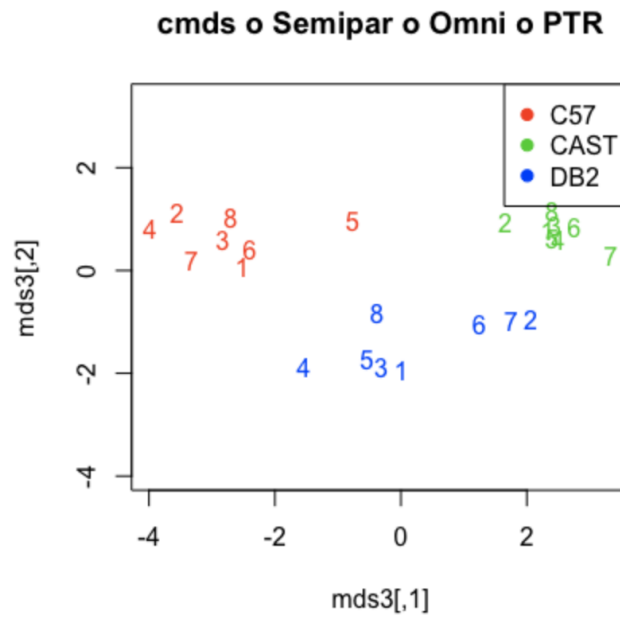


Figure 5.11: CMDS Embedding of the Dissimilarity Matrix.

Bibliography

- L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, pages 211–230, 2003.
- A. Y. Alfakih, A. Khandani, and H. Wolkowicz. Solving Euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12(1):13–30, Jan 1999. ISSN 1573-2894. doi: 10.1023/A:1008655427845. URL <https://doi.org/10.1023/A:1008655427845>.
- J.B. Angelelli, A. Baudot, C. Brun, and A. Guénoche. Two local dissimilarity measures for weighted graphs with application to protein interaction networks. *Advances in Data Analysis and Classification*, 2(1):3–16, 2008.
- A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, Feb 2016. ISSN 0976-8378. doi: 10.1007/s13171-015-0071-x. URL <https://doi.org/10.1007/s13171-015-0071-x>.
- M. Bakonyi and C. Johnson. The Euclidian distance matrix completion problem.

BIBLIOGRAPHY

- SIAM Journal on Matrix Analysis and Applications*, 16(2):646–654, 1995. doi: 10.1137/S0895479893249757. URL <https://doi.org/10.1137/S0895479893249757>.
- M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.
- I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, 2005.
- L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- H. Bunke and K. Riesen. Graph classification based on dissimilarity space embedding. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 996–1007, 2008.
- J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n -way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3):283–319, Sep. 1970. ISSN 1860-0980. doi: 10.1007/BF02310791. URL <https://doi.org/10.1007/BF02310791>.
- D.G. Cattrysse and L.N. Van Wassenhove. A survey of algorithms for the generalized assignment problem. *European Journal of Operational Research*, 60(3):260–272, 1992.
- S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015. doi: 10.1214/14-AOS1272.

BIBLIOGRAPHY

- L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219, 2009. doi: 10.1198/jasa.2009.0111. URL <https://doi.org/10.1198/jasa.2009.0111>.
- D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298, 2004.
- M. A. A. Cox and T. F. Cox. *Multidimensional Scaling*, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-33037-0. doi: 10.1007/978-3-540-33037-0_14. URL https://doi.org/10.1007/978-3-540-33037-0_14.
- T. F. Cox and M. A. A. Cox. *Multidimensional scaling*. Chapman and hall/CRC, 2000.
- A. Criminisi and J. Shotton. Manifold forests. In A. Criminisi and J. Shotton, editors, *Decision Forests for Computer Vision and Medical Image Analysis*, chapter 7, pages 79–94. Springer, London, 2013.
- C. Davis and M. Kahan, W. The rotation of eigenvectors by a perturbation III. *SIAM Journal of Numerical Analysis*, 7:1–46, 1970.
- J. de Leeuw. Applications of convex analysis to multidimensional scaling. In *Recent*

BIBLIOGRAPHY

- Developments in Statistics (Proc. European Meeting Statisticians, Grenoble, 1976)*, pages 133–145. North-Holland, Amsterdam, 1977.
- J. de Leeuw and W. Heiser. Theory of multidimensional scaling. In P.R. Krishnaiah and L. Kanal, editors, *Handbook of Statistics II*, pages 285–316. North Holland Publishing Company, Amsterdam, The Netherlands, 1982.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition (Stochastic Modelling and Applied Probability)*. Springer, corrected edition edition, February 1997.
- J. Fan, Q. Sun, W. X. Zhou, and Z. Zhu. Principal component analysis for big data, January 2018. arXiv:1801.01602.
- D.E. Fishkind, S. Adali, and C.E. Priebe. Seeded graph matching. *arXiv preprint, arXiv:1209.0367*, 2012.
- M.R. Garey and D.S. Johnson. *Computers and intractability: A guide to the theory of NP-completeness*. W.H. Freeman, 1979.
- J. Glaunès, A. Qiu, M. I. Miller, and L. Younes. Large deformation diffeomorphic metric curve mapping. *International Journal of Computer Vision*, 80(3):317–336, 2008. ISSN 0920-5691. URL <http://dx.doi.org/10.1007/s11263-008-0141-9>.
- Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

BIBLIOGRAPHY

- A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
- J. E. Jackson. *A User’s Guide to Principal Components*. Wiley & Sons, New York, 1991.
- A. Javanmard and A. Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, Jun 2013. ISSN 1615-3383. doi: 10.1007/s10208-012-9129-5. URL <https://doi.org/10.1007/s10208-012-9129-5>.
- I. Kaltenmark, B. Charlier, and N. Charon. A general framework for curve and surface comparison and registration with oriented varifolds. *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- K. Levin, A. Athreya, M. Tang, V. Lyzinski, and C. E. Priebe. A central limit theorem for an omnibus embedding of random dot product graphs, 05 2017. arXiv:1705.09355.
- A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
- V. Lyzinski, D.E. Fishkind, and C.E. Priebe. Seeded graph matching for correlated Erdos-Renyi graphs. *arXiv preprint, arXiv:1304.7844*, 2013.

BIBLIOGRAPHY

- V. Lyzinski, Y. Park, C. E Priebe, and M. W. Trosset. Fast embedding for jofc using the raw stress criterion. *Journal of Computational and Graphical Statistics*, 2016.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- J. A. T. Machado and M. E. Mata. Analysis of world economic variables using multidimensional scaling. *PLOS ONE*, 10(3):1–17, 03 2015. doi: 10.1371/journal.pone.0121277. URL <https://doi.org/10.1371/journal.pone.0121277>.
- M. W. Mahoney and P. Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, pages pnas–0803205106, 2009.
- L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- M. Muja and D. G Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(11):2227–2240, 2014.
- S. Oh, A. Montanari, and A. Karbasi. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5, Jan 2010. doi: 10.1109/ITWKSPS.2010.5503144.

BIBLIOGRAPHY

- N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal. Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal Processing Magazine*, 22(4):54–69, July 2005. ISSN 1053-5888. doi: 10.1109/MSP.2005.1458287.
- K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific Publishing Company Inc, Singapore, 2005.
- E. Peterfreund and M. Gavish. Multidimensional Scaling of Noisy High Dimensional Data, January 2018. arXiv:1801.10229.
- R. Prevedel, Y.-G. Yoon, M. Pak, N. Wetzstein, G. Kato, S. Schrodell, T. Raskar, R. Zimmer, M. Boyden, and A. Vaziri. Simultaneous wholeanimal 3d imaging of neuronal activity using light-field microscopy. *Nature Methods*, 2014.
- C.E. Priebe, D.J. Marchette, Z. Ma, and S. Adali. Manifold matching: joint optimization of fidelity and commensurability. *Brazilian Journal of Probability and Statistics*, 27(3):377400, 2013.
- B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In

BIBLIOGRAPHY

- International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.
- B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. doi: 10.1162/089976698300017467. URL <https://doi.org/10.1162/089976698300017467>.
- T. Shi and S. Horvath. Unsupervised learning with random forest predictors. *J. Comput. Graph. Stat.*, 15(1):118–138, March 2006.
- D.B. Shmoys and É. Tardos. An approximation algorithm for the generalized assignment problem. *Mathematical Programming*, 62(1):461–474, 1993.
- A. Singer. A remark on global positioning from local distances. *Proceedings of the National Academy of Sciences*, 105(28):9507–9511, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0709842104. URL <http://www.pnas.org/content/105/28/9507>.
- I. Spence and D. W. Domoney. Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika*, 39(4):469–490, Dec 1974. ISSN 1860-0980. doi: 10.1007/BF02291669. URL <https://doi.org/10.1007/BF02291669>.
- M. J. Taghizadeh. Theoretical analysis of euclidean distance matrix completion for ad hoc microphone array calibration. Idiap-RR Idiap-RR-20-2014, Idiap, 11 2014.
- M. Tang, Y. Park, and C. E. Priebe. Out-of-sample extension for latent position graphs. *arXiv preprint, arXiv:1305.4893*, 2013.

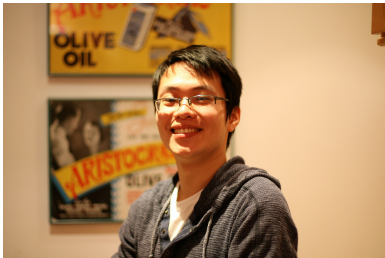
BIBLIOGRAPHY

- A. Tasissa and R. Lai. Exact reconstruction of Euclidean distance geometry problem using low-rank matrix completion. *CoRR*, abs/1804.04310, 2018.
- J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2319. URL <http://science.sciencemag.org/content/290/5500/2319>.
- T. M. Tomita, M. Maggioni, and J. T. Vogelstein. Randomer forests. *arXiv preprint arXiv:1506.03410*, 2015.
- W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.
- J. T. Vogelstein, Y. Park, T. Ohyama, R. A. Kerr, J. W. Truman, C. E. Priebe, and M. Zlatic. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science*, 344(6182):386–392, 2014. ISSN 0036-8075. doi: 10.1126/science.1250298. URL <http://science.sciencemag.org/content/344/6182/386>.
- K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by

BIBLIOGRAPHY

- semidefinite programming. *International journal of computer vision*, 70(1):77–90, 2006.
- L. Yen, M. Saerens, A. Mantrach, and M. Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *in Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–793.
- Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:351–323, 2015.
- L. Zhang, G. Wahba, and M. Yuan. Distance shrinkage and Euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):849–867, 2016. doi: 10.1111/rssb.12138. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12138>.
- M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis*, 51(2):918 – 930, 2006.

Vita



Gongkai (Percy) Li received the B.S. degree in mathematics from the Lehigh University in 2014, the M.S.E. degree in Applied Mathematics and Statistics from the Johns Hopkins University in 2017 and enrolled in the Applied Mathematics and Statistics Ph.D. program at Johns Hopkins University in 2014. He has received the Teaching Fellow Recognition Award.